HZ BOOKS 华章教育

CENGAGE Learning

STATISTICS FOR BUSINESS AND ECONOMICS

12th Edition

商务与经济统计

（原书第12版）

[美] 戴维 R. 安德森（David R. Anderson）辛辛那提大学
丹尼斯 J. 斯威尼（Dennis J. Sweeney）辛辛那提大学
托马斯 A. 威廉斯（Thomas A. Williams）罗切斯特理工学院 著
杰弗里 D. 卡姆（Jeffrey D. Camm）辛辛那提大学
詹姆斯 J. 科克伦（James J. Cochran）路易斯安那理工大学

张建华 王健 冯燕奇 等译

附光盘



机械工业出版社
China Machine Press

STATISTICS FOR BUSINESS AND ECONOMICS

12th Edition

商务与经济统计

（原书第12版）

戴维 R. 安德森（David R. Anderson）辛辛那提大学
丹尼斯 J. 斯威尼（Dennis J. Sweeney）辛辛那提大学
[美] 托马斯 A. 威廉斯（Thomas A. Williams）罗切斯特理工学院 著
杰弗里 D. 卡姆（Jeffrey D. Camm）辛辛那提大学
詹姆斯 J. 科克伦（James J. Cochran）路易斯安那理工大学

张建华 王健 冯燕奇 等译

机械工业出版社

图书在版编目（CIP）数据

商务与经济统计（原书第 12 版）/（美）安德森（Anderson, D. R.）等著；张建华等译．—北京：机械工业出版社，2015.8
（华章教材经典译丛）
书名原文：Statistics for Business & Economics
ISBN 978-7-111-51273-8

I. 商…　II. ①安…　②张…　III. ①商业统计－教材　②经济统计－教材　IV. ① F712.3　② F222

中国版本图书馆 CIP 数据核字（2015）第 200387 号

本书版权登记号：图字：01-2014-0610

David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran.
Statistics for Business & Economics, 12th Edition

本书是美国辛辛那提大学的安德森、斯威尼教授和罗切斯特理工学院的威廉斯教授再度合作的结晶。第 12 版在保留了以前版本的叙述风格和可读性的基础上，对内容进行了一些修订，对个别章节做了更为合理的调整，并更新了一定数量的习题。应用性强是本书的最大特色。作者精心设计了"方法""应用"和"补充练习"三种题型，并设计了起提示、总结和建议作用的"注释和评论"，这些都体现出本书的实用性特点。

本书既可作为统计相关专业本科生、研究生和 MBA 的教材，也可作为从事经济分析工作专业人士的参考读物。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）

责任编辑：左　萌	责任校对：董纪丽
印　　刷：北京诚信伟业印刷有限公司	版　　次：2015 年 9 月第 1 版第 1 次印刷
开　　本：214mm × 275mm　1/16	印　　张：33.25
书　　号：ISBN 978-7-111-51273-8 ISBN 978-7-89405-830-0（光盘）	定　　价：109.00 元（附光盘）

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379210　88361066　　投稿热线：（010）88379007
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjg@hzbook.com

译者序

THE TRANSLATOR'S WORDS

统计学是研究不确定性现象数量规律性的方法论科学，在众多的专业、学科领域中，都起着重要的作用。应用统计学的概念与方法来处理商务与经济中的各种问题，就产生了商务与经济统计学，它是解决商务与经济中各种问题的有力工具之一。

戴维 R. 安德森等五位教授合著的《商务与经济统计》是一本很有特色的教材，我们曾将该书的第7版、第9版、第10版和第11版译成中文介绍给我国的读者，并得到众多读者的认可与积极的反应。2014年《商务与经济统计》的第12版问世了，我们愿以最快的速度再次将新版的中文译本奉献给读者。

新版的《商务与经济统计》保留了以前版本的叙述风格与可读性，其最大特色是应用性强。本书以大量数据为基础，介绍各种统计方法在实际中的应用，每一种统计方法的介绍都联系一个案例，并配有大量的例题和练习。新版的实际数据基本来自《华尔街日报》《今日美国》《财富》等报刊，这使得学生不但对这些统计信息产生兴趣，而且还学习了统计方法及其应用。

本书的另一特色是通俗易懂。本书是为应用者准备的，全书避免了烦琐的数学推导，采用深入浅出、循序渐进的方法系统地介绍了统计学的知识。叙述严谨，基础坚实，实例与图表丰富，易于读者理解与掌握。

本书在每章附录中给出了利用 Minitab，Excel 和 StatTools 进行各种统计分析的程序步骤，并在书后的附录中详细介绍了 Excel 2010，使得学生能够很容易地利用 Minitab，Excel 和 StatTools 完成各种统计分析的运算。

本书安排了大量练习题，用来帮助学生更好地理解书中讲述的内容。其中带有“☆”号的练习题为自测题，以评价学生对书中介绍的各种统计方法的掌握程度。为方便读者自学，在附录C中给出了部分习题的详细解答。

作为教材，本书有较多的应用层次，既可用作研究生、MBA 和本科生的教材，也可供从事工商行政管理和经济分析的各类人员参考。读者可根据时间和需要，有选择地学习有关内容。

为了忠实于原著，我们在翻译过程中没有做任何改动。考虑到我国读者的习惯，我们将原书页边注释改为了页下注。

本书的译者如下：

第1~3章：张建华、张丽华

第4~9章：王健

第10章：聂巧平、叶光

第11~12章：张建华

第13~14章：冯燕奇、刘东红

第15~16章：冯蕾、张坤

第17~20章：张建华、邹华

全书由张建华负责整理和通校，黄河、陈光、王海花、杨莉、杜静、刘伟、周威至、谭欢、刘渊泽、马鑫、黄岩和李超然也参加了本书的部分翻译和校对工作，王公恕通读了全书译稿，并提出了宝贵的意见。

本书在翻译过程中得到了机械工业出版社有关领导的指导帮助，特别要感谢本书的编辑左萌、杨晓莉，他们严谨的工作作风、尽职尽责的工作态度促使本书能在最短的时间内与读者见面。

囿于译者的水平，译文中的疏漏和差错在所难免，恳请读者批评指正。

译者

2014年8月于南开大学

前 言

PREFACE

本书是《商务与经济统计》第12版，本版我们欢迎两位著名的学者：辛辛那提大学的杰弗里 D. 卡姆和路易斯安那理工大学的詹姆斯 J. 科克伦，加入我们的作者团队。杰弗里和詹姆斯二人在统计与商务分析领域都是有造诣的教师、研究员和从业人员。詹姆斯是美国统计协会会员。你可以从随后的作者简介部分了解到他们更多的成就。我们相信杰弗里和詹姆斯的加入，将保持和提高本书想达到的效果。

本书的目的是向学生，主要是向那些工商管理和经济学研究领域的学生概念性地介绍统计学及其各种应用方面的知识。本书是以应用为导向，并考虑到非数学专业人员的需要而编写的，所需要的必备数学基础为代数知识。

数据分析与统计方法的应用是本书的结构及讲述内容中一个不可或缺的部分。每种方法的讨论与发展都通过应用呈现出来，并运用统计结果进行问题的决策和解答。

尽管本书以应用为导向，但我们还是谨慎地给出了合理的方法推导过程，并且运用了所涉及的论题通常会使用的符号。因此，同学们将会发现，本书为学习高级统计学打下了坚实的基础。附录中包括指导进一步学习的参考书目。

本书向学生介绍 Minitab 16 与 Microsoft® Excel 2010 软件包，并强调在统计分析应用中计算机软件包的作用。之所以介绍 Minitab，是因为在教学和统计实践中它都是最主要的统计软件包之一；Excel 虽然不是统计软件包，但它的广泛实用性和应用性，对于学生了解这个软件包的统计性能很重要。Minitab 和 Excel 的步骤在各章附录中会给出，使得在更多强调计算机运用的课堂中，教师可以灵活使用。StatTools 是由 Palisades 公司开发的一个商用 Excel 插件，对 Excel 用户扩展了统计选项的范围。在本书第1章的附录中，我们演示了如何下载和安装 StatTools。大部分章节都有一个演示使用 StatTools 完成统计程序所需步骤的附录。我们将 StatTools 变成可选的，因此对那些讲授仅仅使用 Excel 中可利用的标准统计工具的教师来说，可以这样做。

第12版的变化

我们对《商务与经济统计》以前版本得到的认可与积极的反应表示感谢。因此，在这次新版修订时，我们保留了以前版本的叙述风格与可读性。全书有许多变化旨在提高其教学效果。新版中的显著变化汇总如下。

内容修订

- **描述统计学——第2章和第3章** 我们加入了数据可视化、最佳实践等许多新资料，特别修订了这两章。第2章增加了包括复合条形图和结构条形图的新资料以及数据可视化和创建有效图形显示的最佳实践这一新的小节。第3章在位置的度量一节中增加了对几何平均数的叙述。几何平均数在金融投资的增长率、年百分比等计算中有许多应用。第3章还新增了一节：数据仪表板和如何加入汇总统计以增强其效果。
- **离散型概率分布——第5章** 为了更好地解释概率分布的作用以及第4章中提到的概率分配的资料如何用于建立离散型概率分布，我们对这一章的基础材料进行了修订。我们指出采用分配概率的相对频数法建立经验离散概率分布。应许多用户的要求，我们增加了包括二元离散分布和在金融上的应用这一新的小节（5.4节）。我们展示了如何使用这些分布，构建金融资产组合和分析。
- **多个比率的比较、独立性检验及拟合优度检验——第12章** 这一章做了很大修改。我们增加了三个或多

个总体比率相等性的检验的新的一节，这一节包括一个进行总体比率所有配对之间多重比较检验的程序。为了阐明检验涉及的两个分类变量的独立性，独立性检验这一节被重新改写，修订了包括 Minitab，Excel 和 StatTools 的分步指令。

- **新案例** 本版我们增加了 8 个新案例，使得书中的案例总数达到 31 个。在第 2 章和第 3 章增加了 3 个新的描述统计案例。在有关回归的第 14 章、第 15 章和第 16 章中增加了 5 个新案例。这些案例为学生提供了分析较大数据集并以分析结果为基础准备管理报告的机会。
- **“实践中的统计”的新应用** 每一章都以一篇“实践中的统计”的文章开始，这些文章描述了该章将要介绍的统计方法的应用。本版中新的“实践中的统计”是第 2 章描述的数据仪表板和辛辛那提动物园数据可视化的使用。我们在第 4 章还增加了一个新的“实践中的统计”，它描述 NASA 团队如何利用概率帮助营救由于矿井塌方而受困的 33 名智利矿工。
- **以实际数据为基础的新例题与练习** 我们继续努力用最近的实际数据和统计信息资料来更新本书的例题和练习。本版增加了近 180 道新例题和练习，这些例题和练习是以实际数据和原始资料为基础的。利用《华尔街日报》《今日美国》《巴伦周刊》以及一系列其他资料来源，我们进行了实际研究，以说明并创建练习来演示商务与经济统计的多种应用。我们相信，实际数据的使用可使更多的学生对统计资料产生兴趣，并使学生既学习统计方法，又学习其应用。本书第 12 版包含近 350 道以实际数据为基础的例题和练习。

特色与教学

作者安德森、斯威尼、威廉斯、卡姆和科克伦继续保留了以前版本中的许多特色。

方法练习与应用练习

每节后的练习分成两部分，即“方法”和“应用”。方法练习要求学生利用公式进行必要的计算。应用练习要求学生利用书中的实际资料。这样，学生首先把注意力集中在“基本问题”的计算上，然后再转向精巧的统计应用与解释上。

自测题

某些练习作为自测题（本书用“☆”标注）出现，它们的完整解答在附录 C 中给出，学生们可试着做自测题并核对答案，以便评估你对书中各章讲述的概念的理解程度。

页下注解与注释

关键点以及对学生提供额外知识的页下注解是本书的一个重要特色。设计这些出现在页下的注解，是为了强调和提高对本书介绍的术语和概念的理解。

在许多节的末尾提供了“注释”，以帮助学生更深入地了解统计方法及其应用。“注释”中包括一些注意事项，如方法的局限性、对应用的建议、对其他方法的简要描述及其他事项。

本书附带的数据文件

本书所附光盘有大量可供我们使用的数据文件。[⊖] 这些数据集既可以在 Minitab 中，也可以在 Excel 格式中使用。所有案例的数据集和较大练习的数据集都放在光盘上。

⊖ 本书配套光盘中除了提到的数据文件，还有各章的单数习题、各章附录。

感谢

我们要感谢编审的工作，他们提供的意见和建议有助于继续完善本书。他们是：

AbouEl-Makarim Aboueissa, University of Southern Maine

Kathleen Arano
Fort Hays State University

Musa Ayar
Uw-baraboo/Sauk County

Kathleen Burke
SUNY Cortland

YC Chang
University of Notre Dame

David Chen
Rosemont College and Saint Joseph's University

Margaret E. Cochran
Northwestern State University of Louisiana

Thomas A. Dahlstrom
Eastern University

Anne Drougas
Dominican University

Fesseha Gebremikael
Strayer University/Calhoun Community College

Malcolm C. Gold
University of Wisconsin—Marshfield/Wood County

Joel Goldstein
Western Connecticut State University

Jim Grant
Lewis & Clark College

Reidar Hagtvedt
University of Alberta School of Business

Clifford B. Hawley
West Virginia University

Vance A. Hughey
Western Nevada College

Tony Hunnicutt
Ouachita Technical College

Stacey M. Jones
Albers School of Business and Economics, Seattle University

Dukpa Kim
University of Virginia

Rajaram Krishnan
Earlham College

Robert J. Lemke
Lake Forest College

Philip J. Mizzi
Arizona State University

Mehdi Mohaghegh
Norwich University

Mihail Motzev
Walla Walla University

Somnath Mukhopadhyay
The University of Texas at El Paso

Kenneth E. Murphy
Chapman University

Ogbonnaya John Nwoha
Grambling State University

Claudiney Pereira
Tulane University

J. G. Pitt
University of Toronto

Scott A. Redenius
Brandeis University

Sandra Robertson
Thomas Nelson Community College

Sunil Sapra
California State University, Los Angeles

Kyle Vann Scott
Snead State Community College

Rodney E. Stanley
Tennessee State University

Jennifer Strehler
Oakton Community College

Ronald Stunda
Valdosta State University

Cindy van Es
Cornell University

Jennifer VanGilder
Ursinus College

Jacqueline Wroughton
Northern Kentucky University

Dmitry Yarushkin
Grand View University

David Zimmer
Western Kentucky University

对本书和前几版的构思提供有益意见和建议的许多同事和朋友，我们继续表示感谢。他们是：

Mohammad Ahmadi
University of Tennessee at Chattanooga

Lari Arjomand
Clayton College and State University

Robert Balough
Clarion University

Philip Boudreaux
University of Louisiana

Mike Bourke
Houston Baptist University

James Brannon
University of Wisconsin—Oshkosh

John Bryant
University of Pittsburgh

Peter Bryant
University of Colorado

Terri L. Byczkowski
University of Cincinnati

Robert Carver
Stonehill College

Richard Claycombe
McDaniel College

Robert Cochran
University of Wyoming

Robert Collins
Marquette University

David W. Cravens
Texas Christian University

Tom Dahlstrom
Eastern College

Gopal Dorai
William Patterson University

Nicholas Farnum
California State University—Fullerton

Donald Gren
Salt Lake Community College

Paul Guy
California State University—Chico

Clifford Hawley
West Virginia University

Jim Hightower
California State University, Fullerton

Alan Humphrey
University of Rhode Island

Ann Hussein
Philadelphia College of Textiles and Science

C. Thomas Innis
University of Cincinnati

Ben Isselhardt
Rochester Institute of Technology

Jeffery Jarrett
University of Rhode Island

Ronald Klimberg
St. Joseph's University

David A. Kravitz
George Mason University

David Krueger
St. Cloud State University

John Leschke
University of Virginia

Martin S. Levy
University of Cincinnati

John S. Loucks
St. Edward's University

David Lucking-Reiley
Vanderbilt University

Bala Maniam
Sam Houston State University

Don Marx
University of Alaska, Anchorage

Tom McCullough
University of California—Berkeley

Ronald W. Michener
University of Virginia

Glenn Milligan
Ohio State University

Mitchell Muesham
Sam Houston State University

Roger Myerson
Northwestern University

Richard O'Connell
Miami University of Ohio

Alan Olinsky
Bryant College

Ceyhun Ozgur
Valparaiso University

Tom Pray
Rochester Institute of Technology

Harold Rahmlow
St. Joseph's University

H. V. Ramakrishna
Penn State University at Great Valley

Tom Ryan
Case Western Reserve University

Bill Seaver
University of Tennessee

Alan Smith
Robert Morris College

Willbann Terpening
Gonzaga University

Ted Tsukahara
St. Mary's College of California

Hroki Tsurumi
Rutgers University

David Tufte
University of New Orleans

Victor Ukpolo
Austin Peay State University

Ebenge Usip
Youngstown State University

Cindy Van Es
Cornell University

Jack Vaughn
University of Texas-El Paso

Andrew Welki
John Carroll University

Ari Wijetunga
Morehead State University

J. E. Willis
Louisiana State University

Mustafa Yilmaz
Northeastern University

Gary Yoshimoto
St. Cloud State University

Yan Yu
University of Cincinnati

Charles Zimmerman
Robert Morris College

我们感谢为本书提供“实践中的统计”特色文章的工商业界的合作者，我们通过每一篇文章的来源附加认识了他们中的每个人。我们还要感谢高级策划编辑 Charles McCormick Jr.，初级编辑 Maggie Kubale，内容规划经理 Tamborah Moore，MPS Limited 的项目经理 Lynn Lustberg，媒体编辑 Chris Valentine 以及在本书筹备期间给予支持和编辑建议的 Cengage South-Western 的所有人。

戴维 R. 安德森
丹尼斯 J. 斯威尼
托马斯 A. 威廉斯
杰弗里 D. 卡姆
詹姆斯 J. 科克伦

作者简介

ABOUT THE AUTHOR

戴维 R. 安德森（David R. Anderson）

戴维 R. 安德森是辛辛那提大学工商管理学院数量分析系教授。他出生在北达科他州大福克斯市，在普度大学获得学士、硕士和博士学位。安德森教授担任数量分析与运作管理系主任，并担任工商管理学院的副院长。此外，他还是学院首届教学大纲的协调人。

在辛辛那提大学，安德森教授不但为商科专业的学生讲授基础统计学，而且还讲授研究生水平的回归分析、多元分析和管理科学课程。他也在华盛顿特区的美国劳工部讲授统计学课程。他因在教学上和对学生组织服务方面的突出成就而荣获提名与奖励。

安德森教授已在统计学、管理科学、线性规划以及生产与运作管理领域与他人合作出版了10部著作。他是一位活跃在抽样和统计方法领域的咨询顾问。

丹尼斯 J. 斯威尼（Dennis J. Sweeney）

丹尼斯 J. 斯威尼是辛辛那提大学数量分析系教授和生产力提高中心主任。他出生在衣阿华州得梅因市，在德雷克大学获得工商管理学士学位，在印第安纳大学获得工商管理硕士和工商管理博士学位，并成为NDEA会员。斯威尼教授曾在宝洁公司管理科学小组工作，并在杜克大学作了一年的客座教授。斯威尼教授担任辛辛那提大学数量分析系主任和工商管理学院的副院长。

斯威尼教授已在管理科学与统计学领域发表和出版了30多篇论文和专著。国家科学基金、IBM公司、宝洁公司、美国联合百货（Federated Department Stores）、美国克罗格公司（Kroger）和辛辛那提天然气和电气公司等都曾对他的研究给予资助，这些研究的成果在《管理科学》《运筹学》、*Mathematical Programming*、《决策科学》等杂志上发表。

斯威尼教授在统计学、管理科学、线性规划、生产与运作管理等领域已与他人合作出版了10部专著。

托马斯 A. 威廉斯（Thomas A. Williams）

托马斯 A. 威廉斯是罗切斯特理工学院商学院的管理科学教授。他出生在纽约州埃尔迈拉市，在克拉克森大学获得学士学位，在伦斯勒工学院完成研究生学业并获得硕士和博士学位。

在进入罗切斯特理工学院商学院之前，威廉斯教授在辛辛那提大学工商管理学院从事了7年教学工作，他在那里制定了信息系统专业的本科教学计划，并且担任协调人。在罗切斯特理工学院，他是决策科学系的第一任主席。他不但讲授本科生的管理科学与统计学课程，而且还讲授研究生的回归与决策分析课程。

威廉斯教授在管理科学、统计学、生产与运作管理和数学领域与他人合作出版了11部专著。他为《财富》500强中多家公司提供咨询服务，从数据分析的使用到大型回归模型的开发，都在他的工作范围之内。

杰弗里 D. 卡姆（Jeffrey D. Camm）

杰弗里 D. 卡姆是辛辛那提大学数量分析教授，运营、商业分析与信息系统系主任，卡尔H. 林德纳商学院商

业研究院研究员。他出生在俄亥俄州辛辛那提市，在泽维尔大学获得学士学位，在克莱姆森大学获得博士学位。从1984年起他在辛辛那提大学工作，是斯坦福大学的访问学者和达特茅斯学院商学院工商管理客座教授。

卡姆博士在运营管理中优化问题应用领域发表了30多篇论文。他的研究成果发表在《科学》《管理科学》《运筹学》、*Interfaces*和其他专业杂志上。在辛辛那提大学工作期间，他被任命为教学优秀的Dornoff成员，并在2006年因运筹学实践教学获得了INFORMS奖。他是宣扬实践的坚定信徒。作为运筹学顾问，他服务于多家公司和政府机构。2005~2010年，他担任*Interfaces*总编辑，目前是INFORMS教育学报编委。

詹姆斯 J. 科克伦（James J. Cochran）

詹姆斯J. 科克伦是路易斯安那理工大学拉斯顿银行授衔的数量分析教授。他出生在俄亥俄州的代顿市，在莱特州立大学获得学士、硕士和工商管理硕士学位，在辛辛那提大学获得博士学位。从2000年起在路易斯安那理工大学工作，是斯坦福大学、塔尔卡大学和南非大学的访问学者。

科克伦教授在运筹学和统计方法的发展和应用方面发表了超过24篇论文。他的研究发表在《管理科学》《美国统计》《统计通讯——理论与方法》《运营研究（欧洲版）》，*Journal of Combinatorial Optimization*和其他专业杂志上。在2008年他因运筹学实践教学获得了INFORMS奖，在2010年获Mu Sigma Rho统计教育奖。科克伦教授2005年入选国际统计学会，2010年成为美国统计协会会员。他以提高对实际问题应用质量的方法，强烈主张运筹学和统计学教育的有效性。科克伦教授在乌拉圭的蒙得维的亚、南非的开普敦、哥伦比亚的卡塔赫纳、印度的斋浦尔、阿根廷的布宜诺斯艾利斯、肯尼亚的内罗毕组织和主持教学效果研讨班。作为运筹学顾问，他服务于多家公司和非营利组织。目前他是INFORMS教育学报总编辑及*Interfaces*，*The Journal of the Chilean of Operations Research*及*ORION*的编委。



数理统计及相关课程

课程名称	书号	书名、作者及出版时间	版别	定价
统计学	978-7-111-33687-7	统计学（强森）（2011年）	外版	49
数据、模型与决策	978-7-111-38280-5	数据、模型与决策：管理科学篇（第13版）（安德森）（2012年）	外版	75
数据、模型与决策	978-7-111-49612-0	数据、模型与决策：基于电子表格的建模和案例研究方法（第5版）（希利尔）（2015年）	外版	89
数据、模型与决策	978-7-111-48099-0	数据、模型与决策：基于电子表格的建模和案例研究方法（英文版·第4版）	外版	85
时间序列分析	978-7-111-33864-2	时间序列分析:预测与控制（第4版）（博克斯）（2011年）	外版	59
时间序列分析	978-7-111-38801-2	应用计量经济学：时间序列分析（第3版）（恩德斯）（2012年）	外版	69
商务与经济统计	978-7-111-37641-5	商务与经济统计（第11版）（安德森）（2012年）	外版	108
商务与经济统计	即将出版	商务与经济统计（第12版）（安德森）（2015年）	外版	109
商务与经济统计	978-7-111-24366-3	商务与经济统计（第6版）（纽博尔德）（2008年）	外版	90
商务与经济统计	978-7-111-35029-3	商务与经济统计（英文版·第11版）（安德森）（2011年）	外版	109
商务与经济统计	978-7-111-38666-7	商务与经济统计精要（第6版）（安德森）（2012年）	外版	59
商务与经济统计	978-7-111-40479-8	商务与经济统计精要（英文版·第6版）（安德森）（2012年）	外版	69
经济决策模型	978-7-111-26846-8	经济决策的概率模型（迈尔森）（2009年）	外版	48
计量经济学学习指导	978-7-111-31370-0	经济计量学精要（第4版）习题集（古扎拉蒂）（2010年）	外版	29
计量经济学	978-7-111-30817-1	经济计量学精要（第4版）（古扎拉蒂）（2010年）	外版	49
计量经济学	978-7-111-35537-3	应用计量经济学（第6版）（施图德蒙德）（2011年）	外版	59
运筹学	978-7-111-44029-1	运筹学（第3版）（精品课）（熊伟）（2014年）	本版	35
运筹学	978-7-111-44298-1	运筹学（李峰）（2014年）	本版	39
统计学学习指导	978-7-111-22168-5	应用统计学习指导 （精品课）（孙炎）（2007年）	本版	19
统计学	978-7-111-48630-5	统计学（第2版）（张兆丰）（2014年）	本版	35
统计学	978-7-111-47889-8	统计学（第4版）（精品课）（李金昌）（“十二五”普通高等教育本科国家级规划教材）（2014年）	本版	39
统计学	978-7-111-21720-9	统计学（精品课）（郑珍远）（2007年）	本版	32
统计学	978-7-111-42075-0	统计学（卢小广）（2013年）	本版	35
统计学	978-7-111-42504-5	统计学（向蓉美）（2013年）	本版	39
统计学	978-7-111-31321-2	统计学（曾五一）（2010年）	本版	35
统计学	978-7-111-45966-8	统计学原理（宫春子）（2014年）	本版	35
统计学	978-7-111-29041-4	应用统计基础（精品课）（曾艳英）（2009年）	本版	38
统计学	978-7-111-47018-2	应用统计学（第2版）（精品课）（“十二五”职业教育国家规划教材）（孙炎）（2014年）	本版	35
统计学	978-7-111-44677-4	应用统计学（精品课）（谢忠秋）（2014年）	本版	35
统计、计量分析软件	978-7-111-20747-4	Eviews使用指南与案例（张晓峒）（2007年）	本版	35
数量经济学	978-7-111-26575-7	应用数量经济学（“十一五”国家级规划教材）（张晓峒）（2009年）	本版	45
数据、模型与决策	即将出版	数据、模型与决策（梁樑）（2015年）	本版	40
计量经济学	978-7-111-48459-2	计量经济学（第2版）（李宝仁）（2014年）	本版	39
计量经济学	978-7-111-42076-7	计量经济学基础（ 张兆丰 ）（2013年）	本版	35
计量经济学	即将出版	计量经济学及其应用（第2版）（杜江）（2015年）	本版	39
计量经济学	978-7-111-29842-7	计量经济学及其应用（杜江）（2010年）	本版	29
高级运筹学	978-7-111-24349-6	高级运筹学（马良）（2008年）	本版	30
概率论和数理统计	978-7-111-26974-8	应用概率统计（彭美云）（2009年）	本版	27
概率论和数理统计	978-7-111-28975-3	应用概率统计学习指导与习题选解（彭美云）（2009年）	本版	18

经济学









课程名称	书号	书名、作者及出版时间	版别	定价
预测理论与方法	978-7-111-38130-3	经济预测基础教程（第4版）（迪博尔德）（2012年）	外版	59
国际经济学	978-7-111-34033-1	国际经济学（第8版）（赫斯特德）（2011年）	外版	62
国际经济学	978-7-111-35280-8	国际经济学（英文版·第8版）（赫斯特德）（2011年）	外版	69
国际经济学	978-7-111-49021-0	国际经济学：国际金融分册（第8版）（阿普尔亚德）（2015年）	外版	39
国际经济学	978-7-111-47783-9	国际经济学：国际贸易分册（第8版）（阿普尔亚德）（2014年）	外版	59
管理经济学	978-7-111-48424-0	管理经济学（第11版）（莫瑞斯）（2014年）	外版	75
管理经济学	978-7-111-39670-3	管理经济学（第12版）（麦圭根）（2012年）	外版	89
管理经济学	978-7-111-41881-8	管理经济学（第12版）（中国版）（麦圭根）（2013年）	外版	55
管理经济学	978-7-111-31879-8	管理经济学（第9版）（中国版）（托马斯、刘延平）（2010年）	外版	48
管理经济学	978-7-111-26569-6	管理经济学（英文版·第11版）（麦圭根）（2009年）	外版	68
管理经济学	978-7-111-48567-4	管理经济学（英文版·第11版）（莫瑞斯）（2014年）	外版	89
发展经济学	978-7-111-45840-1	发展经济学（第11版）（托达罗）（2014年）	外版	89
财政学	即将出版	财政学（第4版）（格鲁伯）（2015年）	外版	79
经贸英语	978-7-111-40348-7	经贸英语阅读教程（第3版）（潘忠）（2012年）	本版	25
经贸英语	978-7-111-44666-8	商务英语阅读（沈素萍）（2013年）	本版	35
经贸英语	978-7-111-46218-7	商务英语阅读教程（杨伶俐）（2014年）	本版	29
经济法	978-7-111-47546-0	经济法（第2版）（葛恒云）（2014年）	本版	35
经济法	978-7-111-44659-0	经济法（第2版）（郭懿美）（2014年）	本版	39
经济法	978-7-111-41779-8	经济法（刘大洪）（2013年）	本版	39
经济法	978-7-111-13974-4	经济法基础与实务（黄瑞）（2008年）	本版	32
国际经济学	978-7-111-25578-9	国际经济学（精品课）（赵英军）（2009年）	本版	32
国际经济关系学	978-7-111-27371-4	国际经济关系学概论（周林）（2009年）	本版	30
管理经济学	978-7-111-39608-6	管理经济学（毛蕴诗）（"十二五"普通高等教育本科国家级规划教材）（2012年）	本版	45
产业经济学	978-7-111-49568-0	产业经济学（刘志彪）（2015年）	本版	39
财政学	978-7-111-27276-2	财政学（第2版）（李友元）（2009年）	本版	36
财经应用文写作	978-7-111-42715-5	财经应用文写作（刘常宝）（2013年）	本版	30
博弈论	978-7-111-30394-7	博弈论及其应用（李帮义）（2010年）	本版	25



华章教材经典译丛（清明上河图）系列









课程名称	书号	书名、作者及出版时间	定价
财务会计	即将出版	财务会计：概念、方法与应用（第14版）（威尔）（2015年）	95
财务会计	978-7-111-39244-6	财务会计教程（第10版）（亨格瑞）（2012年）	79
财务管理（公司理财）学习指导	978-7-111-32466-9	公司理财（第8版）习题集（汉森）（2010年）	42
财务管理（公司理财）	978-7-111-36751-2	公司理财（第9版）（罗斯）（2012年）	88
财务管理（公司理财）	978-7-111-47887-4	公司理财（精要版）（第10版）（罗斯）（2014年）	75
电子商务	978-7-111-45187-7	电子商务：管理与社会网络的视角（第7版）（特班）（2014年）	79
战略管理	978-7-111-39138-8	战略管理:概念与案例（第8版）（希尔）（2012年）	69
战略管理	978-7-111-43844-1	战略管理：获取持续的竞争优势（第4版）（巴尼）（2013年）	69
商业伦理学	978-7-111-37513-5	企业伦理学（第7版）（乔治）（2012年）	79
领导学	978-7-111-47356-5	领导学（全球版·第8版）（尤克尔）（2014年）	65
管理学	978-7-111-46255-2	管理学（诺里亚）（2014年）	69
管理学	978-7-111-41449-0	管理学：原理与实践（第8版）（罗宾斯）（2013年）	59
管理学	即将出版	管理学：原理与实践（第9版）（罗宾斯）（2015年）	59
管理技能	978-7-111-37591-3	管理技能开发（第8版）（惠顿）（2012年）	98
创业管理	即将出版	百森创业教学法：基于实践的视角（奈克）（2015年）	49
创业管理	978-7-111-40258-9	公司创新与创业（第3版）（库拉特科）（2012年）	49
项目管理	978-7-111-39774-8	项目管理：基于团队的方法（布朗）（2012年）	49
数据、模型与决策	978-7-111-49612-0	数据、模型与决策：基于电子表格的建模和案例研究方法（第5版）（希利尔）（2015年）	89
管理会计	978-7-111-39512-6	管理会计教程（第15版）（亨格瑞）（2012年）	88
投资银行学	978-7-111-41476-6	投资银行、对冲基金和私募股权投资（斯托厄尔）（2013年）	99
金融中介学	978-7-111-43694-2	金融市场与金融机构（第7版）（米什金）（2013年）	99
金融学（货币银行学）指导或案例	978-7-111-44311-7	货币金融学（第2版）学习指导（米什金）（2013年）	45
金融学（货币银行学）	978-7-111-34261-8	货币金融学（第2版）（米什金）（2011年）	75
金融市场学	978-7-111-26674-7	金融市场学（第10版）（罗斯）（2009年）	79
金融工程学习指导	978-7-111-30014-4	期权、期货及其他衍生产品习题集（第7版）（赫尔）（2010年）	42
金融工程	978-7-111-48437-0	期权、期货及其他衍生产品（第9版）（赫尔）（2014年）	109
（证券）投资学学习指导	978-7-111-42662-2	投资学习题集（第9版）（博迪）（2013年）	49
（证券）投资学	978-7-111-39028-2	投资学（第9版）（博迪）（2012年）	98
（证券）投资学	978-7-111-44485-5	投资学（第9版）（专业版）（博迪）（2013年）	199
中级宏观经济学	978-7-111-43155-8	宏观经济学（第5版）（布兰查德）（2013年）	75
西方经济学学习指导	978-7-111-33099-8	哈伯德《经济学》学习指南（第3版）（斯卡希尔）（2011年）	45
西方经济学学习指导	978-7-111-31352-6	经济学精要（精要版）（第4版）学习指南（拉什）（2010年）	39
西方经济学（微观）	978-7-111-32767-7	经济学（微观）（第3版）（哈伯德）（2011年）	59
西方经济学（微观）	978-7-111-42810-7	经济学（微观部分）（第2版）（斯通）（2013年）	55
西方经济学（宏观）	978-7-111-32768-4	经济学（宏观）（第3版）（哈伯德）（2011年）	49
西方经济学（宏观）	978-7-111-42849-7	经济学（宏观部分）（第2版）（斯通）（2013年）	49
西方经济学	978-7-111-28088-0	经济学：私人与公共选择（第12版）（格瓦特尼）（2009年）	78
西方经济学	978-7-111-27481-0	经济学原理（精要版）（第4版）（帕金）（2009年）	62
商务与经济统计	978-7-111-37641-5	商务与经济统计（第11版）（安德森）（2012年）	108
商务与经济统计	即将出版	商务与经济统计（第12版）（安德森）（2015年）	109
财政学	即将出版	财政学（第4版）（格鲁伯）（2015年）	79
组织行为学	978-7-111-44814-3	组织行为学精要（第12版）（罗宾斯）（2014年）	45
人力资源管理	978-7-111-40189-6	人力资源管理（亚洲版·第2版）（德斯勒）（2012年）	65
消费者行为学	978-7-111-47509-5	消费者行为学（第12版）（霍金斯）（2014年）	79
市场营销学（营销管理）	978-7-111-43017-9	市场营销学（第11版）（阿姆斯特朗、科特勒）（2013年）	75
市场营销学（营销管理）	978-7-111-43202-9	市场营销原理（亚洲版·第3版）（科特勒）（2013年）	79
服务营销学	978-7-111-48495-0	服务营销（第6版）（泽丝曼尔）（2014年）	75
供应链（物流）管理	978-7-111-45565-3	供应链物流管理（第4版）（鲍尔索克斯）（2014年）	59
管理信息系统	978-7-111-34151-2	管理信息系统（第11版）（劳顿）（2011年）	55

目录

CONTENTS

⊖ 请参见本书所配光盘。

第1章

数据与统计资料

CHAPTER 1

实践中的统计

彭博商业周刊 [⊖]
纽约州，纽约市

《彭博商业周刊》（*Bloomberg Businessweek*）是世界上拥有最广泛读者的商业类杂志，其全球发行量超过100万份。该杂志在全球的145个记者站中有1 700多名忠于职守的记者和编辑，主要登载各类有关全球经济和贸易活动的饶有兴趣的文章。除了刊登当前热门话题的特色文章外，还有一些关于国际贸易、经济分析、信息处理和科技进展等常设栏目。特色文章和常设栏目传递的信息能帮助读者及时透视时事风云变幻，以便评估其对商务和经济状况产生的影响。

几乎每一期《彭博商业周刊》（更名前叫《商业周刊》）都发表一篇有关当前热门话题的深度报告，通常这些深度报告包含统计论据和概述，以帮助读者了解商务和经济信息。例如，2011年3月3日那一期的封面故事，讨论了工商业逐渐把其主要业务处理向云计算转移；2011年5月30日那一期，有一篇濒临崩溃的美国邮政服务业的报告；2011年8月1日那一期，发表了一篇为什么债务危机比想象的更为严重的文章。另外，每一期《彭博商业周刊》都发布关于经济状况的统计资料，包括产品指数、股票价格、共同基金和利息率等。

《彭博商业周刊》还利用统计报告和统计信息来帮助管理其自身的经营。例如，一份关于订阅者的年度综述报告可以帮助公司了解订阅者的人口特征、阅读习惯、购物偏好和生活方式等。《彭博商业周刊》的管理人员利用从调查中得出的统计结论更好地为订阅者和广告客户提供服务。最近对《彭博商业周刊》北美订阅者的调查数据表明，90%的《彭博商业周刊》订阅者在家使用计算机，并且64%的订阅者因工作需要而购买了计算机。这一统计结果提醒《彭博商业周刊》的管理者：订阅者会对个人计算机发展趋势方面的文章感兴趣，而且，调查的结论也会引起潜在的广告客户的兴趣。订阅者中在家使用计算机和因工作使用计算机有如此之高的百分比，将会刺激个人计算机厂商考虑在《彭博商业周刊》上做广告。

本章我们讨论可用于统计分析的数据类型，并说明如何取得这些数据。我们将介绍描述统计和统计推断，它们是将数据转为有意义的且易于解释的统计信息的方法。

我们经常在报纸和杂志上的文章中看到下列各种报道：

- 美国劳工部报道失业率降至过去3年的最低点8.2%（《华盛顿邮报》，2012年4月6日）。
- 每年每一位美国人平均消费23.2夸脱的冰淇淋、牛奶冻、果汁牛奶冻、冰冻甜食和其他大量生产的冷冻乳制品（makeicecream. com website，2012年4月2日）。
- 度假屋销售价格的中位数为121 300美元（@CNNMoney，2012年3月29日）。
- 位于田纳西州Pigeon Forge的多来坞主题公园的野鹰宽翼过山车最大时速为每小时98千米（USA Today website，2012年4月5日）。
- 钉板方式的社交图片分享网站Pinterest的注册用户数量从1月中旬到2月中旬增加了85%（CNBC，2012年3月29日）。
- Pew研究中心报道美国新娘初婚年龄的中位数创纪录高达26.5岁（Significance，2012年2月）。
- 2011年第4季度加拿大人上网时间平均达到45小时（CBC News，2012年3月2日）。
- 美联储报道每个人的平均信用卡债务为5 204美元（PRWeb website，2012年4月5日）。

前面报道中的数据事实（8.2%，23.2，121 300美元，61，85%，26.5，45，5 204美元）称为**统计资料**（statistics）。在这种用法中，术语统计资料是指数据值事实，如平均数、中位数、百分数和最大值等，它们可以帮助我们了解各种商务和经济状况。然而，你将看到，统计学的研究领域或对象的内涵更加丰富。广义上讲，统

⊖ 作者感谢为“实践中的统计”提供了案例的调研管理人员 Charlene Trentham。

计学是搜集、分析、表述和解释数据的科学与艺术。特别是在商务和经济活动中，搜集、分析、表述和解释数据旨在帮助管理者和决策者更好地理解商务和经济环境的变化，并作出科学、正确的决策。在本书中，我们强调的是统计学在商务和经济决策方面的用途。

本章以统计学在商务和经济中的一些应用实例为开端，在第1.2节我们定义了数据这一术语的含义，并给出了数据集的概念。在这一节中，还介绍了变量和观测值等一些重要的术语，讨论了数量型数据和分类型数据之间的区别，并且说明了截面数据和时间序列数据的应用。在第1.3节中，我们探讨了如何从现有的资料来源中获得数据，或通过调查和设计的实验研究获得新数据。对如今在数据获取中起着重要作用的互联网，也给予了高度的重视。在第1.4节和第1.5节中，论述了数据在进行描述统计和统计推断中的用途。在本章的最后3节还给出了计算机在统计分析中的角色，简单介绍了数据挖掘，并且讨论了统计实践的道德准则。本章末尾的附录中对Excel的插件StatTools做了介绍，它扩展了Microsoft Excel用户的统计选项功能。

1.1 统计学在商务和经济中的应用

如今在贸易和经济全球化的背景下，人们可以获取大量的统计信息。最成功的管理者和决策者是那些能够理解信息并有效利用信息的人。本节我们举例说明统计学在商务和经济中的应用。

1.1.1 会计

会计师事务所在对其客户进行审计时需要使用统计抽样程序。例如，假设一个事务所想确定列示在客户资产负债表上的应收账款金额是否真实地反映了应收账款的实际金额。通常应收账款的数量是如此之大，以致查看和验证每一账户将花费大量的时间和费用。在这种情况下，一般的做法是：审计人员从账户中选择一个子集作为样本，在查看样本账户的准确性后，审计师得出有关列示在客户资产负债表上的应收账款金额是不是可以接受的结论。

1.1.2 财务

财务顾问们利用各种各样的统计信息指导投资。在股票市场中，财务顾问们综合了包括市盈率和股息等方面的财务数据，通过对比单只股票和股票市场平均状况的信息，就可以得出某一只股票是否具有投资价格的结论。例如，《华尔街日报》（2012年3月19日）报道说，标准普尔500公司的平均股息收益率是2.2%，微软的股息收益是2.42%。这种情况下，关于股息收益率的统计信息表明，微软的股息收益率比标准普尔500公司的平均收益率高。这一信息和其他有关微软的信息将帮助财务顾问作出是买、是卖还是持股的建议。

1.1.3 市场营销

零售结账柜台的电子扫描仪正用于搜集各种市场调研用的数据。例如，数据供应商ACNielsen公司和Information Resources公司，从商店购买POS扫描数据，经过加工处理，作出统计汇总后再出售给制造商。制造商为取得这些扫描数据，每一类产品要花费数十万美元。制造商也购买特价销售和利用店内陈列品等促销活动的数据及统计研究报告。产品品牌经理可以查看扫描资料和促销活动统计资料，从而能更好地理解促销活动和销售额之间的关系。这样的分析对制定各种产品未来的市场营销战略大有裨益。

1.1.4 生产

由于当今进入了重视质量的时代，因此质量管理是统计学在生产中的一项重要应用。各种统计质量管理图用于监测生产过程的产出。特别地，用$\bar{x}$控制图可以监测平均产出。例如，假定有一台340克的软饮料罐装机，定期地从产品中选择一些罐装饮料作为样本，计算出样本罐装量的平均值。这一平均值或$\bar{x}$的值标在一张$\bar{x}$控制图上。当该数值位于控制上限以上时，则表明产品罐装量过高；当该数值位于控制下限以下时，则表明产品罐装量不足；当$\bar{x}$的值位于控制图的控制上限和控制下限之间时，这一过程称为处于“受控”状态，并且只要$\bar{x}$的值落

在控制图的控制上限和控制下限之间，就允许连续生产。更贴切的解释是，$\bar{x}$ 控制图能帮助确定何时必须调整和修正生产过程。

1.1.5 经济

人们经常要求经济学家对未来的经济或某一方面的发展作出预测。他们在进行预测时需要用到各种统计信息。例如，在预测通货膨胀率时，经济学家利用诸如生产者价格指数、失业率和制造业开工率等指标的统计信息。这些统计指标往往要输入到预测通货膨胀率的计算机预测模型中。

本节介绍的统计学的这些应用是本书的组成部分。这些例子可使我们对统计学的广泛应用有一个大致的了解。为了补充这些例子，我们请求商务和经济领域的专家提供每章开篇的"实践中的统计"，以便引出每章所要讲解的内容。"实践中的统计" 充分显示统计学在各种各样的商务和经济问题中的重要性。

1.1.6 信息系统

信息系统管理员负责维护某个组织的计算机网络日复一日地运行。大量的统计信息帮助管理员评价计算机网络［包括局域网（LANs）、广域网（WANs）、网络段、内部网和其他数据通信信息］的表现。统计诸如系统中的平均用户数、系统中组件下载次数的比例和日常宽带使用的比例都是统计信息的例子。这些统计信息可以帮助系统管理员更好地了解和管理计算机网络。

1.2 数据

数据（data）是为了描述和解释所搜集、分析、汇总的事实和数字。将用于特定研究而搜集的所有数据称为研究的**数据集**（data set）。表 1-1 是一个包含参加世界贸易组织的 60 个国家信息的数据集。世界贸易组织支持国际贸易自由化，并提供了解决贸易争端的场所。

1.2.1 个体、变量和观测值

个体（element）是指搜集数据的实体。在表 1-1 的数据集中，每一个国家是一个个体，国家或个体的名称列在表 1-1 中的第 1 列。有 60 个国家，数据集中就有 60 个个体。

变量（variable）是个体中所感兴趣的那些特征。表 1-1 的数据集中有下列 5 个变量：

- WTO 身份：在世界贸易组织中国家的成员身份，它可以是成员或观察员。
- 人均 GDP（美元）：国家的总产出除以该国的总人口数，它通常用于比较国家的经济生产率。
- 贸易逆差（1 000 美元）：国家的进口总额与出口总额之差。
- 惠誉评级：由惠誉国际组织评价的国家主权信用评级 [⊖]；信用评级从最高的 AAA 到最低的 F，并且可以用 + 或 - 来微调。
- 惠誉评级展望：未来 2 年内信用评级可能变动的方向性指标，展望可以是正面、稳定或负面。

表 1-1 世界贸易组织 60 个国家的数据集

国家	WTO 身份	人均 GDP（美元）	贸易逆差（1 000 美元）	惠誉评级	惠誉评级展望
亚美尼亚	成员	5 400	2 673 359	BB -	稳定
澳大利亚	成员	40 800	-33 304 157	AAA	稳定
奥地利	成员	41 700	12 796 558	AAA	稳定
阿塞拜疆	观察员	5 400	-16 747 320	BBB -	正面
巴林	成员	27 300	3 102 665	BBB	稳定
比利时	成员	37 600	-14 930 833	AA +	负面

⊖ 惠誉国际组织是国际公认的三大统计评价机构之一，由美国证券交易监督委员会设计。另外两大机构是标准普尔和穆迪投资者服务公司。

（续）

国家	WTO 身份	人均 GDP（美元）	贸易逆差（1 000 美元）	惠誉评级	惠誉评级展望
巴西	成员	11 600	−29 796 166	BBB	稳定
保加利亚	成员	13 500	4 049 237	BBB−	正面
加拿大	成员	40 300	−1611 380	AAA	稳定
佛得角	成员	4 000	874 459	B+	稳定
智利	成员	16 100	−14 558 218	A+	稳定
中国	成员	8 400	−156 705 311	A+	稳定
哥伦比亚	成员	10 100	−1 561 199	BBB−	稳定
哥斯达黎加	成员	11 500	5 807 509	BB+	稳定
克罗地亚	成员	18 300	8 108 103	BBB−	负面
塞浦路斯	成员	29 100	6 623 337	BBB	负面
捷克共和国	成员	25 900	−10 749 467	A+	正面
丹麦	成员	40 200	−15 057 343	AAA	稳定
厄瓜多尔	成员	8 300	1 993 819	B−	稳定
埃及	成员	6 500	28 486 933	BB	负面
萨尔瓦多	成员	7 600	5 019 363	BB	稳定
爱沙尼亚	成员	20 200	802 234	A+	稳定
法国	成员	35 000	118 841 542	AAA	稳定
格鲁吉亚	成员	5 400	4 398 153	B+	正面
德国	成员	37 900	−213 367 685	AAA	稳定
匈牙利	成员	19 600	−9 421 301	BBB−	负面
冰岛	成员	38 000	−504 939	BB+	稳定
爱尔兰	成员	39 500	−59 093 323	BBB+	负面
以色列	成员	31 000	6 722 291	A	稳定
意大利	成员	30 100	33 568 668	A+	负面
日本	成员	34 300	31 675 424	AA	负面
哈萨克斯坦	观察员	13 000	−33 220 437	BBB	正面
肯尼亚	成员	1 700	9 174 198	B+	稳定
拉脱维亚	成员	15 400	2 448 053	BBB−	正面
黎巴嫩	观察员	15 600	13 715 550	B	稳定
立陶宛	成员	18 700	3 359 641	BBB	正面
马来西亚	成员	15 600	−39 420 064	A−	稳定
墨西哥	成员	15 100	1 288 122	BBB	稳定
秘鲁	成员	10 000	−7 888 993	BBB	稳定
菲律宾	成员	4 100	15 667 209	BB+	稳定
波兰	成员	20 100	19 552 976	A−	稳定
葡萄牙	成员	23 200	21 060 508	BBB−	负面
韩国	成员	31 700	−37 509 141	A+	稳定
罗马尼亚	成员	12 300	13 323 709	BBB−	稳定
俄罗斯	观察员	16 700	−151 400 000	BBB	正面
卢旺达	成员	1 300	939 222	B	稳定
塞尔维亚	观察员	10 700	8 275 693	BB−	稳定
塞舌尔	观察员	24 700	666 026	B	稳定
新加坡	成员	59 900	−27 110 421	AAA	稳定
斯洛伐克	成员	23 400	−2 110 626	A+	稳定
斯洛文尼亚	成员	29 100	2 310 617	AA−	负面
南非	成员	11 000	3 321 801	BBB+	稳定
瑞典	成员	40 600	−10 903 251	AAA	稳定
瑞士	成员	43 400	−27 197 873	AAA	稳定
泰国	成员	9 700	2 049 669	BBB	稳定
土耳其	成员	14 600	71 612 947	BB+	正面
英国	成员	35 900	162 316 831	AAA	负面
乌拉圭	成员	15 400	2 662 628	BB	正面
美国	成员	48 100	784 438 559	AAA	稳定
赞比亚	成员	1 600	−1 805 198	B+	稳定

在一项研究中，对每个个体的每一变量收集测量值，从而得到了数据。对某一特定个体得到的测量值集合称为一个**观测值**（observation）。如表 1-1 所示，我们看到第一个观测值（亚美尼亚）的测量值集合是成员、5 400、2 673 359、BB - 和稳定。第二个观测值（澳大利亚）的测量值集合是成员、40 800、- 33 304 157、AAA 和稳定。60 个个体的数据集有 60 个观测值。

1.2.2 测量尺度

搜集数据需要按下列几种测量尺度之一来度量：名义尺度、顺序尺度、间隔尺度和比率尺度。测量尺度决定了数据中蕴涵的信息量，并表明最适合的数据汇总和统计分析方法。

当一个变量的数据包含了用来识别个体属性的标记或名称时，测量尺度称为**名义尺度**（nominal scale）。例如，表 1-1 中的数据，我们看到 WTO 身份变量的测量尺度是名义尺度，因为数据"成员"和"观测员"是用来识别国家身份类型的标记。在测量尺度是名义尺度的情形中，可以使用数值代码及非数字的标记。例如，为了便于数据搜集，并做好将数据录入计算机数据库的准备，对 WTO 身份变量我们可以使用数值代码：用 1 代表世界贸易组织中的成员国；用 2 代表观测员国。即使数据是数值型的，测量尺度仍是名义尺度的。

如果数据具有名义数据的性质，并且数据的顺序或等级的意义明确，那么这种变量的测量尺度是**顺序尺度**（ordinal scale）。如表 1-1 中所示，惠誉评级的测量尺度是顺序尺度，因为从 AAA 到 F 的评级分类可以按从最佳的信用评级 AAA 到最差的信用评级 F 排序。评级字母提供了类似名义数据的标记，但是数据还可以根据信用评级排序，这使得测量尺度是顺序的。顺序尺度也可以用数值代码，例如你在学校中的班级排名。

如果数据具有顺序数据的所有性质，并且可以按某一固定度量单位表示数值间的间隔，则这种变量的测量尺度是**间隔尺度**（interval scale）。间隔数据永远是数值型的。学生能力测验（SAT）的分数是间隔尺度数据的一个例子。例如，3 个学生的 SAT 分数分别为 620、550 和 470，它们能够按最好到最差进行排序。另外，分数之差是有一定意义的。比如，学生 1 的分数比学生 2 的分数多 70 分（ = 620 - 550），学生 2 的分数比学生 3 的分数多 80 分（ = 550 - 470）。

如果数据具有间隔数据的所有性质，并且两个数值之比是有意义的，则这种变量的测量尺度是**比率尺度**（ratio scale）。像距离、高度、重量和时间等变量都用比率尺度来度量。比率尺度需要有一个零值，变量取零值时表示什么也不存在。例如，让我们来考虑汽车的成本，零值意味着汽车没有成本或是免费的。另外，如果我们比较成本为 30 000 美元的汽车与成本为 15 000 美元的汽车，比率值 30 000/15 000 = 2，表示第一辆汽车的成本是第二辆汽车成本的 2 倍。

1.2.3 分类型数据和数量型数据

数据还可以进一步划分为分类型和数值型。归属于某一类别的数据称为**分类型数据**（categorical data），分类型数据既可以用名义尺度度量也可以用顺序尺度度量。用于表示大小或多少的数值称为**数量型数据**（quantitative data）。数值型数据既可以用间隔尺度度量也可以用比率尺度度量。

分类变量（categorical variable）是用分类型数据表示的变量，**数量变量**（quantitative variable）是用数量型数据表示的变量。统计分析方法是否适合一个特定变量，取决于变量是分类变量还是数量变量[㊀]。如果变量是分类变量，则统计分析方法极其有限。我们通过记录每一类别中观测值的数目，或计算每一类别中观测值的比例来汇总分类型数据。但是，即使分类型数据用数值代码表示，对其进行加、减、乘和除等数学运算也是没有意义的。第 2.1 节将介绍汇总分类型数据的方法。

另一方面，对数量变量进行数学运算，可以得到有意义的结果。例如，对于数量变量，可以先求和，然后除以观测值的个数，可以计算出平均数。这个平均数通常是有意义的并且易于解释。一般地，当数据是数值型时，有更多的统计方法可供选择。第 2.2 节和第 3 章会介绍汇总数量型数据的方法。

㊀ 适合汇总数据的统计方法取决于数据是分类型数据还是数值型数据。

1.2.4 截面数据和时间序列数据

为了便于统计分析，有必要对截面数据和时间序列数据进行区分。**截面数据**（cross-sectional data）是在相同或近似相同的同一时点上搜集的数据。表 1-1 中数据是截面数据，因为它们描述了 60 个世界贸易组织国家的 5 个变量在同一时点上的情况。**时间序列数据**（time series data）是在几个时期内搜集的数据。例如，图 1-1 的时间序列是 2007～2012 年美国常规普通汽油每加仑的平均价格曲线。我们注意到，汽油价格在 2008 年夏季达到顶点后，2008 年下半年汽油价格急剧下降。从 2008 年后，每加仑汽油平均价格重新开始稳步攀升，到 2012 年再次接近历史新高。



图 1-1 美国常规普通汽油每加仑的平均价格

资料来源：Energy Information Administration，U. S. Department of Energy（美国能源部能源情报署），2012 年 3 月。

在商务与经济出版物中经常可以看到时间序列数据的图形，这些图形可以帮助分析家们了解过去发生的情况，确定现象随时间变动的趋势，推测时间序列的未来水平。时间序列数据的图形有各种形态，如图 1-2所示。稍加研究，这些图形通常易于理解和解释。例如，图 1-2a 是 2002～2012 年道琼斯工业平均价格指数的图形。在 2002 年 4 月普通股票市场指数接近 10 000 点，在随后的 5 年中股指不断攀升，到 2007 年 10 月超过 14 000 点。然而，在 2007 年创历史新高后，股指急速下降。到 2009 年 3 月，糟糕的经济条件迫使道琼斯工业平均价格指数重新回到了 7 000 点的水平。对投资者来说，这是一个令人惊慌、失去信心的时期。然而，到 2009 年年末，股指有所复苏，达到 10 000 点。到 2012 年年初，股指稳步攀升，已超过 13 000 点。

图 1-2b 是 2005～2011 年麦当劳公司的净收益图形，2008 年和 2009 年的经济衰退，实际上对麦当劳公司而言是受益的，期间其净收益创了历史新高。麦当劳公司的净收益的增长表明公司的繁荣是由于在经济衰落期人们减少了去相对昂贵的餐厅就餐，转而寻求麦当劳提供更便宜的替代食品。在 2010～2011 年，麦当劳公司的净收益继续创历史新高。

图 1-2c 是一年期间佛罗里达州南部旅店入住率的一个时间序列图形。在 2～3 月，当佛罗里达州南部的气候对游客有吸引力时，出现 95%～98% 的极高入住率。事实上，每年的 1～4 月是佛罗里达州南部旅店典型高入住率的季节。另外，8～10 月入住率较低，最低的 50% 入住率出现在 9 月，高温以及飓风季节是这段时间入住率下降的主要原因。

注释和评论

1. 观测值是数据集中每个个体的测量值的集合。因此，观测值的个数总是与个体的个数一致。每个个体的测量值的个数等于变量的个数。因此，数据项的总数是个体的个数乘以变量的个数。
2. 数量型数据可以是离散的也可以是连续的。度量可数事物多少的数量型数据是离散的（例如，5 分钟内接到电话的个数），度量不可数事物的多少的数量型数据是连续的（例如，体重或时间）。

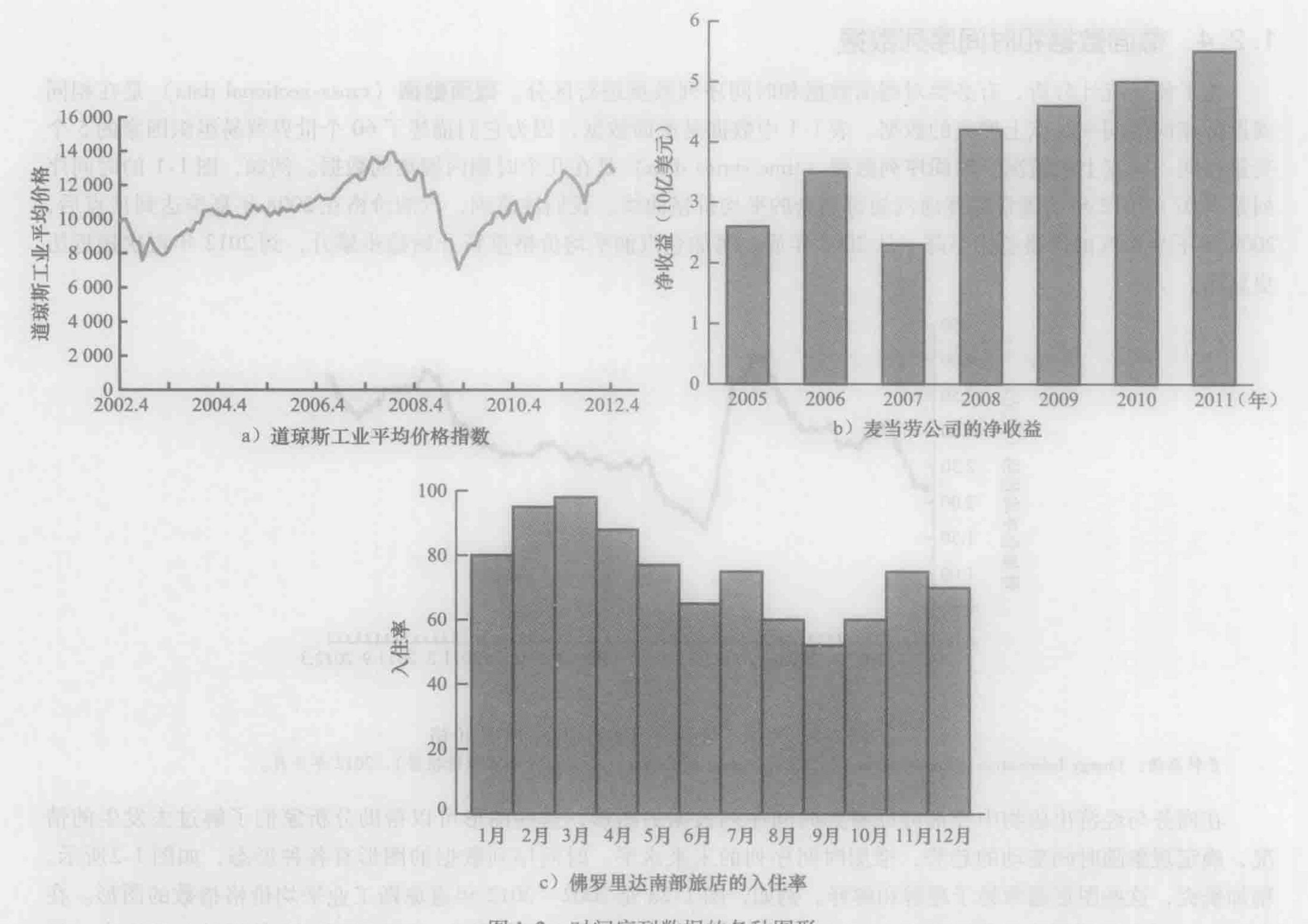


图 1-2 时间序列数据的各种图形

1.3 数据来源

数据可以从现有来源中搜集，或通过调查获得，亦可通过为了得到新数据而设计的实验性研究获得。

1.3.1 现有来源

在某些情况下，为某一特定应用所需要的数据已经存在。公司保持有关其雇员、客户和义务运作等各种数据库。关于雇员工薪、年龄和服务年限的数据通常可以从内部个人记录中获得；关于销售收入、广告支出、配送成本、库存水平以及产量的数据通常可以从其他内部记录中获得；绝大多数公司还保存其客户的详细资料。表 1-2 是一些从公司内部记录得到的常用的数据。

表 1-2 从公司内部记录的可用数据的例子

数据来源	一些典型数据的可用数据
雇员记录	姓名、地址、社会保障号码、工资、休假天数、病假天数和奖金
生产记录	部件或产品号、生产数量、直接人工成本和原材料成本
存货记录	部件或产品号、库存单位数、重复订货水平、经济订货批量和折扣表
销售记录	产品号、销售量、区域销售量和分类客户销售量
信用记录	客户名称、地址、电话号码、信贷额度和应收账款余额
客户概况	年龄、性别、收入水平、家庭规模、地址和偏好

大量的商务和经济数据可以从专门从事搜集和保存数据的机构那里获得。公司可以通过租赁或购买的方式使用这些来自外部的数据。邓白氏、彭博社和道琼斯公司是三家能够为客户提供大量商业数据服务的公司。AC-Nielsen 和 Information Resources 公司也成功地为广告商和制造商提供商务数据搜集和加工服务。

数据也可从各类行业协会和专门营利的机构中获得。美国旅游业协会保存各州与旅游相关的信息，如游客数量和旅游费用等，旅游行业的公司和个人会对这些数据感兴趣。研究生管理咨询委员会保存了关于考试成绩、学生特长和研究生培养计划等的数据。大多数来自这些类型的数据用户使用时需要支付适当的费用。

互联网的持续发展，已使其成为数据和统计信息的一个重要来源。几乎所有的公司都会提供公司的一般信息，如销售额、雇员数、产量、产品价格和产品说明的网站。另外，大多数公司专门在互联网获取有用的市场信息资料。因此，人们能从网上查阅到股票报价、餐馆菜价、工资数据等几乎无限的各种信息。

政府机构是现有数据的另一个重要来源。例如，美国劳工部保存大量的有关就业率、工资率、劳动力规模和工会会员等数据。表 1-3 列示了部分政府机构和它们能提供的一些数据。大部分政府机构也通过网站向社会提供其搜集和加工处理的数据。图 1-3 显示的是美国劳工统计局的主页。

表 1-3　来自选取政府机构的可用数据的例子

政府机构	一些可用的数据
人口普查局	人口数、家庭数和家庭收入的数据
联邦储备委员会	货币供应、信誉记录、汇率和贴现率的数据
管理和预算办公室	财政收入、支出和联邦政府债务的数据
商务部	商务活动数据、分行业总产值、分行业利润水平以及行业增长和下降的数据
劳工统计局	消费支出、小时工资、失业率、安全记录和国际统计资料



图 1-3　美国劳工统计局主页

1.3.2　统计研究

有时一些特殊需要的数据并非总能从现有来源中获得。在这种情况下，常常通过统计研究的方式获得数据。统计研究可划分为实验性的或观测性的两类。

在一项实验性的研究[⊖]中，首先要确定感兴趣的主要变量，然后控制一个或多个其他变量，以便获得它们如何影响主要变量的数据。例如，一家制药公司可能会进行一项实验，以获得一种新药如何影响血压的情况。在研究中，血压是被关注的主要变量，新药的剂量是影响血压的另一个变量，实验希望能找出新药的剂量与血压之间

⊖ 曾经进行过的最大规模的实验性统计研究是 1954 年美国公共卫生署的沙克脊髓灰质炎疫苗实验，该研究从全美小学 1～3 年级的学生中选出了近 200 万名儿童。

的因果关系。为了获得有关新药效果的数据，研究人员选择一些个体组成样本。控制新药的剂量，不同的个体组给予不同的药剂水平，然后搜集每组服药前后的血压数据。实验数据的统计分析将会有助于了解新药如何影响血压。

在非实验性或观测性统计研究[⊖]中，并不是有目的地去控制感兴趣的变量。调查也许是观测性研究中最常用的方法。例如，在一个征求顾客意见的调查中，首先要明确研究的问题，然后设计调查问卷，选择一些人作为样本。一些饭店利用观测研究来获取顾客对饭店的食品质量、服务和就餐环境等方面的数据。佛罗里达州那不勒斯的 Chops City Grill 饭店使用的顾客意见调查表如图 1-4 所示。注意，请顾客根据自己的感受对综合经验、接待问候、管理（餐桌巡视）和整体服务等 12 个变量按优秀、良好、平均、普通和较差的等级顺序打分，得到的数据能帮助 Chops City Grill 的管理人员保持高水平的饭店食品和服务质量。

Chops
CITY GRILL

日期：________ 服务员姓名：________

顾客是我们应予以最优先考虑的。请占用您一点点宝贵时间填写这张调查表，这将使我们更好地为您提供服务。您可将此表交到前台或邮寄给我们，谢谢！

调查项目	优秀	良好	平均	普通	较差
综合经验	❑	❑	❑	❑	❑
接待问候	❑	❑	❑	❑	❑
管理（餐桌巡视）	❑	❑	❑	❑	❑
整体服务	❑	❑	❑	❑	❑
职业素养	❑	❑	❑	❑	❑
菜单知识	❑	❑	❑	❑	❑
服务态度	❑	❑	❑	❑	❑
酒品选择	❑	❑	❑	❑	❑
菜单选择	❑	❑	❑	❑	❑
食品质量	❑	❑	❑	❑	❑
食品展示	❑	❑	❑	❑	❑
消费（美元）	❑	❑	❑	❑	❑

对改进我们的服务您还有何建议？

谢谢，非常感谢您的建议。——Chops City Grill的全体员工

图 1-4 佛罗里达州那不勒斯的 Chops City Grill 饭店的顾客意见问卷调查表

将数据和统计分析结果用于管理决策，管理人员必须清楚获得数据所需要的时间和成本。当必须在一个相对较短的时期内获得数据时，利用现有数据来源是一个比较理想的方法，如果重要的数据不容易从现有来源中获取，则必须考虑获取数据所要占用的时间和成本。在所有的情形中，决策制定者必须意识到统计分析对决策制定过程中的贡献。数据搜集和统计分析所付出的成本是不会超过利用这些信息制定一个好的决策所节省的费用的。

1.3.3 数据采集误差

管理人员任何时候都应牢记：统计研究中的数据是可能产生误差的。使用错误的数据也许会比不使用任何数据更糟糕。每当取得的数据值不等于真值或实际值时，就会产生数据搜集误差（真值或实际值是可以通过一个正确的程序获得），这类错误可能在许多数据搜集方法中发生。例如，采访者可能会发生记录错误，如将 24 岁写为

⊖ 对吸烟者和不吸烟者的研究是观测性研究，因为调查者并不确知或控制那些人会将吸烟，那些人不吸烟。

42 岁，或者被采访者回答问题时可能曲解了问题并作出不正确回答。

实验数据分析人员在搜集和记录数据时要特别小心，以确保不发生错误。可采取一些特殊措施以检查数据的内在一致性。例如，这种措施会提醒分析人员查看一下有关数据的准确性：如一个年龄为 22 岁的人却说有 20 年的工作经验。数据分析人员也应注意查看一些过大或过小的被称为异常值的数值，它们可能是有错误的数据。在第 3 章我们将介绍一些统计学家用来识别异常值的方法。

在数据搜集过程中会经常发生误差。盲目地利用一切可以得到的或不认真地搜集到的数据，将可能产生误差并导致作出错误的决策。因此，按照正确的步骤搜集准确的数据可以确保决策信息的可信度，提高数据的利用价值。

1.4　描述统计

绝大多数的报纸、杂志、公司报告和其他出版物上的统计信息以读者易于理解的方式汇总和披露。将数据以表格、图形或数值形式汇总的统计方法被称为**描述统计**（descriptive statistics）。

再来参见表 1-1 中参加世界贸易组织的 60 个国家的数据集。用描述统计的方法对这一数据集中的信息进行汇总。例如，考虑未来 2 年内国家信用评级可能变动的方向性指标——惠誉评级展望变量，该变量用负面、稳定和正面来记录。表 1-4 给出了每一个成员国惠誉评级展望数据的表格汇总，图 1-5 给出了相同数据的图形汇总，该类图形称为条形图。这些表格和图形汇总一般能使数据更容易解释。我们能很容易地从表 1-4 和图 1-5 中看到，绝大多数信用评级的惠誉展望是稳定，有 65% 的国家具有这个评级。信用评级的负面展望和正面展望相似，具有负面展望的国家（18.3%）比具有正面展望的国家（16.7%）略微多一些。

图 1-6 是被称为直方图的表 1-1 中人均 GDP 这个数量变量数据的图形汇总。从直方图上很容易看出，60 个国家的人均 GDP 的变动范围为 0～60 000 美元，主要集中在 10 000～20 000 美元，只有 1 个国家的人均 GDP 超过 50 000 美元。

表 1-4　60 个国家惠誉信用评级展望类型的频数和百分数

惠誉评级展望	频数	百分数（%）
正面	10	16.7
稳定	39	65.0
负面	11	18.3

除了表格和图形表示方式以外，数值型描述统计也用于汇总数据。最常用的数值型描述统计是平均数或均值。利用表 1-1 中 60 个国家人均 GDP 数据，我们可以将所有 60 个国家的人均 GDP 相加再除以 60，计算出平均数。这样可得到人均 GDP 的平均数为 21 387 美元。这个平均数是数据的中心趋势或中心位置的一个度量。

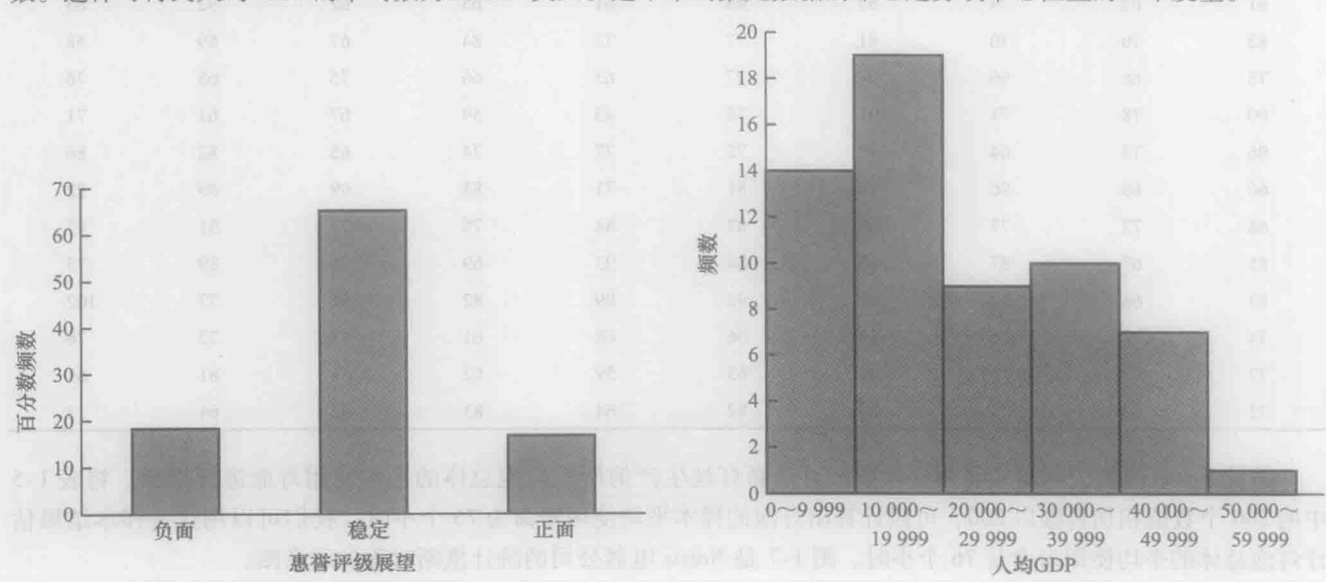


图 1-5　60 个国家惠誉信用评级展望的条形图

图 1-6　60 个国家人均 GDP 的直方图

近年来，人们关注描述统计的拓展和表述，对这类统计方法的兴趣在不断提高。在第2章和第3章中将集中讨论描述统计的表格、图形和数值方法。

1.5 统计推断

在很多情形下，我们需要搜集有关大的个体组（个人、公司、选民、家庭、产品和客户等）的信息。出于对时间、费用和其他因素的考虑，只能从大的个体组的一小部分中搜集数据。在特定研究中，较大的个体组称为**总体**（population），较小的个体组称为**样本**（sample）。我们一般使用下列定义。

> **总体** 总体是在一个特定研究中所有个体组成的集合。
>
> **样本** 样本是总体的一个子集。

搜集总体全部数据的调查过程称为**普查**（census）[⊖]，搜集样本数据的调查过程称为**抽样调查**（sample survey）。统计学的一个主要贡献就是利用样本数据对总体特征进行估计和假设检验。这一过程被称为**统计推断**（statistical inference）。

作为统计推断的一个例子，让我们考察由Norris电器公司所做的一项研究。Norris电器公司生产一种能在各种电器产品上使用的高亮度灯泡。为了提高灯泡的使用寿命，产品设计部门开发出一种使用新型灯丝的灯泡。在这个例子中，总体定义为使用新灯丝生产的所有灯泡。为了估计新灯丝的优点，用生产出的200只新灯丝灯泡组成样本，并进行测试，记录每只灯泡在灯丝被烧断之前使用的小时数，样本数据见表1-5。

表1-5 Norris电器公司200只灯泡的使用寿命 （单位：小时）

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73

假设Norris电器公司希望利用样本数据对用新灯丝生产的所有灯泡总体的平均使用寿命进行推断。将表1-5中的200个数值相加再除以200，可以计算出灯泡的样本平均使用寿命为76个小时。我们可以用这一样本结果估计灯泡总体的平均使用寿命是76个小时。图1-7是Norris电器公司的统计推断过程的示意图。

⊖ 美国政府每10年进行一次人口普查，而市场调查机构每天进行抽样调查。

每当统计学家利用样本对我们感兴趣的总体的特征进行估计时，他们通常都要注明估计的质量或精确度。以Norris电器公司为例，统计学家可能指出，新灯泡总体的平均使用寿命的点估计值是76个小时，边际误差为±4小时。这样，新灯泡总体的平均使用寿命的区间估计是72~80个小时。统计学家也可以指出，在72~80个小时的区间内有多大的把握程度包含的总体平均数。

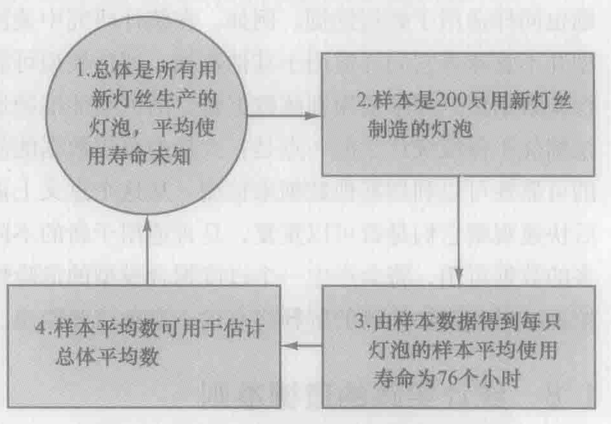


图1-7 Norris电器公司统计推断的过程

1.6 计算机与统计分析

由于涉及大量的数据，因此统计人员经常使用计算机软件进行分析。例如，在Norris电器公司的例子中（见表1-5），如果没有计算机，那么计算200只灯泡的平均使用寿命将相当乏味。为了方便计算机的使用，教材中较大的数据集可以从本书所附光盘获得，下载的数据文件适合于Minitab和Excel格式。本章末的附录还有使用Minitab，Excel和Excel的插件StatTools来完成本章所介绍的统计方法的详细步骤。

1.7 数据挖掘

由于有磁卡片输入机、条形码扫描仪和POS终端，许多机构每天可以获取大量的基础数据。即使一个当地的使用触摸屏显示器来登录订单和处理账单的小饭店，搜集大量的数据也是有意义的。对于大型零售公司，搜集数据的真实值仅仅是概念性的，并且弄清楚如何有效地使用这些数据来提高效率也是一个难题。例如，大型零售商（如沃尔玛）每天获取2 000万~3 000万个交易数据，电信公司（如法国电信和美国电话电报公司）每天有超过3亿个电话记录，Visa每秒处理6 800次支付交易，或每天处理近6亿个交易。储存和管理这些交易数据是一件有意义的事情。

术语数据仓库是指获取、储存和维护数据的过程。计算能力和数据搜集工具可以在几秒内实现储存和检索非常庞大的数据。分析仓库中的数据可以对机构作出新的战略和较高收益的决策。

数据挖掘（data mining）问题是研究从非常大的数据库中开发有用的决策信息的方法。利用统计学、数学和计算机科学的综合方法，分析人员在仓库中“挖掘数据”，并将其转化为有用的信息，取名为数据挖掘。数据挖掘领域的先驱Kurt Thearling博士定义数据挖掘为“从（庞大的）数据库中自动提取预测性的信息”。Thearling博士的定义中的两个关键词是“自动”和“预测性的”。数据挖掘系统是最有效的使用自动程序从数据中提取信息，而其用户仅仅需要使用最一般的甚至是模糊的查询。同时数据挖掘软件使揭示隐藏的预测信息的过程自动化，而这些隐藏的信息过去是需要亲自动手进行分析。

拥有大型客服中心的公司（如零售商、金融机构和通信公司）成为数据挖掘的主要应用者。数据挖掘成果用于帮助零售商（如亚马逊和巴诺）确定已经购买了一件特殊商品的消费者，是否还有可能购买一件或多件相关的商品。然后，当消费者登录公司网站并且购买一件商品时，网站利用突然跳出窗口来促使消费者有可能追加购买额外商品。在另一个应用中，数据挖掘用于确定在某个特殊的购物旅行中消费有可能超过20美元的消费者，然后将这些消费者划分为收到特别电子邮件赠券和定期电子邮件赠券的类型，从而在赠券到期之前促使他们进行下一次购物旅行。

数据挖掘是一门非常依赖于统计方法的技术，这些统计方法有多元回归和逻辑回归等。它创造性地综合了所有这些方法以及包括人工智能和机器学习等计算机科学技术，从而使数据挖掘是有效的。由诸如甲骨文、Teradata和SAS等公司开发的用以实施商用数据挖掘的软件包，需要大量的时间和金钱上的投入。本书介绍的统计概念将有助于理解数据挖掘软件包中使用的统计方法，并使你更好地理解其所产生的统计信息。

由于统计模型在建立数据挖掘预测模型中扮演着重要的角色[⊖]，许多统计学家在建立统计模型中所关注的问题也同样适用于数据挖掘。例如，在统计研究中关注的问题是模型的可靠性，适用于一个特殊样本数据的统计模型并不意味着它同样适用于其他数据。评估模型可靠性的一个常见方法是将样本数据分为两部分：训练数据集和检验数据集。如果使用训练数据建立的模型能准确地预测检验数据的数值，则我们认为这个模型是可靠的。数据挖掘优于传统统计学的一点是：大量的可用数据能使数据挖掘软件划分数据集，因此由训练数据集所建立的模型的可靠性可以利用其他数据来检测。从这个意义上讲，数据集的划分允许用数据挖掘来建立模型和相互关系，然后快速观察它们是否可以重复，是否适用于新的不同的数据。另一方面，对数据挖掘应用的一个忠告是，有这么多的数据可用，将会产生一个过度拟合模型的危险性，在这一点似乎存在误导性的联系和因果关系的结论。数据挖掘结果和附加检测的解释将有助于避免这些隐患。

1.8 统计实践的道德准则

道德行为是我们在一切活动中的追求。由于统计在数据的搜集、分析、表述和说明过程中的重要作用，道德问题继而在统计中产生。在统计研究中，不道德行为形式很多，其中包括不正确的抽样、不恰当的数据分析、误导性的图表绘制、不恰当地使用统计摘要和统计结果的片面解释，等等。

你开始做统计工作时，我们就鼓励你在搜集数据、分析、口头陈述以及在书面报告中，要保持公平、谨慎、客观和中立的态度。另外，作为数据的使用者，你也应该意识到他人在统计过程中不道德行为的可能性。当你在报纸、电视、互联网或其他媒介上见到统计时，要持几分怀疑态度，关注这些统计的来源、目的和客观性。

美国统计协会——美国最高统计和统计工作者的专业组织，发布报告“统计实践的道德准则”[⊜]，旨在帮助统计工作者作出符合道德准则的决定，并交流心得，帮助学生学会如何负责任地进行统计工作。报告中包括 67 条准则、8 个主要领域，分别为：职业精神；对投资者、客户和雇主的责任；对出版物和证明的责任；对研究课题的责任；对研究团队的同事的责任；对其他统计学家或统计从业人员的责任；关于不当行为的指控的责任；机构、个人、代理人的雇主或其他聘用的统计从业人员的客户的责任。

职业精神领域的一条准则解决的问题是多次检测直到获得理想结果的行为。例如，在 1.5 节中我们讨论了 Norris 电器公司的一个 200 只新灯丝高亮度灯泡的样本的统计研究，样本的平均使用寿命为 76 个小时，由此推断出所有新灯丝灯泡的使用寿命。然而，思考一下，由于 Norris 只选取了一个样本，那么另外一个样本可能会提供不同的平均使用寿命，这么想不是没有道理的。

假定 Norris 的管理层希望抽样结果能支持他们声称的新灯泡的平均使用寿命至少是 80 个小时。再假定 Norris 的管理层决定继续研究，不断生产并抽样检测，每次都抽 200 只新灯丝灯泡，直到有一个样本的平均使用寿命达到至少 80 个小时。如果研究重复进行许多次，最终会有一个样本（仅是偶然）给出期望的结果，从而支持 Norris 作出的声明。在这种情形下，消费者将受到误导，认为新产品会比现在的要好。显然，这种行为是不道德的，它代表统计在实践中出现了重大的使用错误。

对出版物和证明领域里有几条道德准则涉及数据的处理问题。例如，一位统计学家需要说明一项研究中考虑的所有数据，并对实际上使用的样本进行解释。Norris 电器公司最初样本中的 200 只灯泡的使用寿命是 76 个小时；这比管理层希望得到的至少 80 个小时的结果要小很多。现在假设在回顾了平均使用寿命为 76 个小时的结果之后，Norris 剔除所有使用寿命小于或等于 70 个小时的观察值，并宣称生产过程中的启动问题造成了这些灯泡的瑕疵。之后，样本中剩余灯泡的平均寿命就成了 82 个小时。你对 Norris 所宣布的灯泡的平均寿命是 82 个小时产

⊖ 在揭示数据中的关系和预测未来结果时，统计方法在数据挖掘中扮演着重要的角色。然而，数据挖掘的全部领域以及统计在数据挖掘中的应用已超出本书的范围。

⊜ American Statistical Association “Ethical Guidelines for Statistical Practice”, 1999.

生过怀疑吗?

如果Norris把使用寿命小于或等于70个小时的灯泡剔除，只是为了提供82个小时的平均使用寿命，那么毫无疑问这种做法就是不道德的。但是，如果这些灯泡真是由于在生产过程中的启动问题造成的瑕疵，因而不应该包含在分析中，进行研究的统计人员应该考虑所有的数据并对抽样的具体操作进行解释。否则，就可能产生误导，对公司和统计人员都构成不道德行为。

在美国统计协会报告的共享价值部分有一条准则提出，统计从业人员应当避免任何对预测结果有倾向性的统计工作。当使用不具有代表性的样本而作出声明时，这种不道德的行为就会发生。例如，在美国的许多地区，饭店内是不允许吸烟的。假如烟草业的说客在允许吸烟的饭店进行调查，目的是估计赞成在饭店内允许吸烟的比例。抽样结果表明90%的受访者同意在饭店内允许吸烟。根据这个抽样结果，说客声称饭店所有就餐者中有90%的人赞成在饭店内允许吸烟。在这个案例中，我们应该指出只在允许吸烟的饭店所做的抽样会歪曲调查结果。如果只报告该研究的这个最后结果，而不熟悉研究的细节（如样本是在允许吸烟的饭店搜集的），读者就会被误导。

美国统计协会报告的涵盖范围很广，道德准则不仅适用于统计工作者，同样对统计信息的使用者也有效。当你从事统计研究工作时，我们建议你读一读这个报告，对道德相关问题有更好的了解，并获得一些背景知识，以确保当你开始在实践中使用统计工作时，你的工作符合道德标准。

小　结

统计是搜集、分析、表述和解释数据的艺术和科学。几乎每个商务或经济专业的大学生都要学习统计学课程。我们以描写商务和经济中典型的统计应用开始本章的论述。

数据是指搜集和分析的事实和数字。我们可以用四种测量尺度来取得一个特定变量的数据，它们是名义尺度、顺序尺度、间隔尺度和比率尺度。当数据用于确认个体属性的标记或名称时，相应变量的测量尺度称为名义尺度。如果数据显示名义数据的性质，并且数据的顺序或等级是有意义的，则测量尺度是顺序尺度。如果数据具有顺序数据的性质，并且可以按某一固定度量单位表示数值间的间隔，则测量尺度是间隔尺度。最后，如果数据具有间隔数据的所有性质，并且两个数值之比是有意义的，则测量尺度是比率尺度。

为了统计分析之目的，数据可划分为分类型数据和数量型数据。分类型数据是用标记或名称来识别每一个体属性。分类型数据既可以用名义尺度度量也可以用顺序尺度度量，既可以是非数值型的也可以是数值型的。数量型数据是表示大小或多少的数值。数量型数据既可以用间隔尺度度量也可以用比率尺度度量。只有当数据是数量型的，普通的算术运算才有意义。因而，适合于数量型数据的统计计算并非总适用于分类型数据。

在第1.4和第1.5节我们介绍了描述统计和统计推断。描述统计是用表格、图形和数值方法来汇总数据。统计推断是利用样本数据估计总体特征并进行假设检验的过程。在本章最后三节，给出了计算机在统计分析中作用的信息，介绍了数据挖掘这个比较新的领域，并且简要描述统计实践的道德准则。

关键术语

statistics　统计学　搜集、分析、表述和解释数据的艺术和科学。

data　数据　是描述和解释所搜集、分析和汇总的事实和数字。

data set　数据集　在特定研究中搜集到的所有数据。

elements　个体　搜集数据的实体。

variable　变量　个体中所感兴趣的特征。

observation　观测值　对某一特定个体搜集的测量值的集合。

nominal scale　名义尺度　是指用数据的标记或名称来识别个体属性的一种变量测量尺度，名义数据既可以是非数值型的，也可以是数值型的。

ordinal scale　顺序尺度　具有名义数据的性质，并能显示数据的顺序或等级有明确意义的一种变量测量尺度。顺序数据既可以是非数值型的，也可以是数值型的。

interval scale 间隔尺度 具有顺序数据的性质，并可以按某一固定度量单位表示数值间的间隔的一种变量测量尺度。间隔数据永远是数值型的。

ratio scale 比率尺度 具有所有间隔数据的性质，并且两个数值之比是有意义的一种变量测量尺度。比率数据永远是数值型的。

categorical data 分类型数据 用于识别每一个体属性的标记或名称。分类型数据既可以用名义尺度度量也可以用顺序尺度度量，既可以是非数值型的也可以是数值型的。

quantitative data 数量型数据 表示事物大小或多少的数值。数量型数据既可以用间隔尺度度量也可以用比率尺度度量。

categorical variable 分类变量 用分类型数据表示的变量。

quantitative variable 数量变量 用数量型数据表示的变量。

cross-sectional data 截面数据 在相同或近似相同的同一时点上搜集的数据。

time series data 时间序列数据 在几个时期内搜集的数据。

descriptive statistics 描述统计 数据的表格、图形和数值汇总的方法。

population 总体 在一个特定研究中所有感兴趣的个体组成的集合。

sample 样本 总体的一个子集。

census 普查 搜集总体全部数据的调查。

sample survey 抽样调查 搜集样本数据的调查。

statistical inference 统计推断 利用样本数据估计或对总体特征进行假设检验的过程。

data mining 数据挖掘 利用统计和计算机科学的方法从非常大的数据库中获取有用信息的过程。

补充练习[㊀]

2. 美国能源部提供各种汽车燃料燃烧效率的信息。表1-6是由10辆汽车组成的一个样本（Fuel Economy website，2008年2月22日）。数据有汽车类型（小型、中型或大型）、发动机的汽缸数、每加仑城市行驶里程（单位：英里）、每加仑[㊁]公路行驶里程以及推荐使用的燃料（柴油、优质汽油或普通汽油）。
   a. 这个数据集中有多少个体？
   b. 这个数据集中有多少变量？
   c. 哪些变量是分类变量，哪些变量是数量变量？
   d. 每个变量使用哪种类型的测量尺度？

4. 表1-7中给出7所大学的数据，其中有捐赠额（单位：10亿美元）和申请者被接受的比例（《今日美国》，2008年2月3日）。每所学校所在的州、校园所处位置以及学校代表队在美国大学生体育总会（NCAA）的分级可以从美国教育统计中心网站（2008年2月22日）上得到。
   a. 这个数据集中有多少个体？
   b. 这个数据集中有多少个变量？
   c. 哪些变量是分类变量，哪些是数量变量？

表1-6 10辆汽车燃料燃烧效率信息

汽车	类型	汽缸数	城市行驶（英里/加仑）	公路行驶（英里/加仑）	燃料
奥迪 A8	大型	12	13	19	优质汽油
宝马 328Xi	小型	6	17	25	优质汽油
凯迪拉克 CTS	中型	6	16	25	普通汽油
克莱斯勒 300	大型	8	13	18	优质汽油
福特福克斯	小型	4	24	33	普通汽油
现代伊兰特	中型	4	25	33	普通汽油
吉普大切诺基	中型	6	17	26	柴油
庞蒂亚克 G6	小型	6	15	22	普通汽油
丰田凯美瑞	中型	4	21	31	普通汽油
大众捷达	小型	5	21	29	普通汽油

㊀ 本书所有章节的习题部分中单数题请参见本书配套光盘。
㊁ 1加仑（美制）≈3.8升。

表1-7 7所大学的数据

学校	州	校园位置	捐赠（10亿美元）	申请者被接受的比例（%）	NCAA的分级
安默特斯学院	马萨诸塞	城镇：边缘	1.7	18	Ⅲ
杜克大学	北卡罗来纳	城市：中型	5.9	21	I-A
哈佛大学	马萨诸塞	城市：中型	34.6	9	I-AA
斯沃斯莫尔学院	宾夕法尼亚	市郊：大型	1.4	18	Ⅲ
宾夕法尼亚大学	宾夕法尼亚	城市：大型	6.6	18	I-AA
威廉姆斯学院	马萨诸塞	城镇：边缘	1.9	18	Ⅲ
耶鲁大学	康涅狄格	城市：中型	22.5	9	I-AA

6.《外交事务》杂志进行了一项调查，以了解其订阅者的简况（《外交事务》，2008年2月23日）。提出的问题如下：

a. 在过去的12个月内您有几天住在旅店？

b. 您在哪儿购书？列出了3个选项：书店、互联网和读书俱乐部

c. 您是否拥有或租赁过豪华车？（是或否）

d. 您多大年龄？

e. 在过去3年的国外旅行中，您的目的地是哪里？列出了7个国外目的地

对每一个问题的数据是分类型的还是数量型的发表意见。

8. 金融时报/哈里斯民意调查对6个欧洲国家和美国的成年人每月进行一次民意在线调查。美国有1 015名成年人参加了1月份的民意调查，其中一个问题是："您如何评价联邦银行在处理金融市场信用问题时的表现？"可能的答案有：优秀、良好、中等、较差和极差（Harris Interactive website，2008年1月）

a. 这项调查的样本容量是多大？

b. 数据是分类型的还是数量型的？

c. 对这一问题，用平均数或百分比数据汇总，哪种方法更合适？

d. 在美国的回答者中，10%的人说联邦银行做得很好，有多少人给出这种答复？

10. 运输统计局综合住房调查是一项年度调查，并作为美国交通运输部的信息来源。在部分调查中，采访了调查者，并询问如下问题："机动车驾驶员在开车时是否可以手持电话接听。"可能的答案有：非常同意、稍微同意、稍微反对和非常反对。对这个问题有44人表示非常同意、130人表示稍微同意、165人表示稍微反对、741人表示非常反对（Bureau of Transportation website，2010年8月）。

a. 这个问题的答案提供的是分类型数据还是数量型数据？

b. 对这一问题的答案，用平均数或百分比数据汇总，哪种方法更合适？

c. 表示非常同意允许机动车驾驶员在开车时可以手持电话接听的答案占多大比例？

d. 该结果预示普遍支持还是反对允许机动车驾驶员在开车时可以手持电话接听？

12. 夏威夷旅游局搜集在夏威夷的游客数据。下列问题是在飞往夏威夷的航班上，向乘客分发的问卷中所征询的16个问题中的一部分。

- 此行是我的第几次夏威夷之旅：第1，2，3，4…次
- 此行的主要理由是：（包括度假、惯例和蜜月等10种）
- 我计划住在：（包括旅店、公寓、亲戚和宿营等11类）
- 在夏威夷逗留的总天数

a. 被研究的总体是什么？

b. 对飞往夏威夷航班的乘客总体，用问卷调查是一个好方法吗？

c. 对4个问题中的每一个所提供的是分类型数据还是数量型数据进行评论。

14. 下面是2007～2010年Hertz，Avis和Dollar这3家提供租车服务的公司的租车数量数据（单位：1 000辆）。

公司	2007	2008	2009	2010
Hertz	327	311	286	290
Dollar	167	140	106	108
Avis	204	220	300	270

a. 绘制2007～2010年每个租车公司租车数量的时间序列图，在同一张图上显示3家公司的时间

序列。

b. 市场占有率最大的租车公司是哪家？对市场占有率如何随时间变动加以评价。

c. 用2010年的租车数据绘制条形图。这张图是依据截面数据还是时间序列数据绘制的？

16. 美国能源部能源信息局提供了2007年1月至2012年3月之间美国常规普通汽油每加仑平均价格的时间序列数据（Energy Information Administration website，2012年4月）。利用互联网获取2012年3月以来的常规普通汽油每加仑的平均价格。

a. 扩展图1-1的时间序列图形。

b. 你如何解释从2012年3月以来常规普通汽油每加仑的平均价格？

c. 时间序列是否继续呈现每加仑的平均价格在夏季上涨？请解释。

18. 关于完税、在线纳税筹划和电子申报服务，舆论研究公司对1 021名成年人（18岁及以上）进行电话随机调查。调查结果表明其中有684人计划使用电子申报他们的税务（CompleteTax Tax Prep Survey，2010年）。

a. 确定一种描述统计量，能用于估计所有使用电子申报的纳税人的百分数。

b. 调查报告说，人们最常用的报税方法是雇佣会计师或职业报税人。如果有60%的调查者使用了这种方式，那么有多少人会雇佣会计师或职业报税人？

c. 人们申报纳税常用的其他方法还有手工填报、使用在线税务服务和使用税务软件。申报纳税方法的数据是分类型数据还是数量型数据？

20.《巴伦周刊》Big Money对131名投资经理进行了一次民意调查，结果如下：

- 43%的经理认为他们自己在股票市场操作极佳。
- 在接下来的12个月里，股票的平均预期回报率为11.2%。
- 21%的经理认为医疗健康类股票极有可能是接下来12个月中股票市场的主导板块。
- 当问及预计科技股和电信股估计大约需要多长时间才恢复上涨时，他们的回答都是2.5年。

a. 列举两种描述统计。

b. 对接下来12个月里投资经理总体的股票平均预期回报率作出推断。

c. 对科技股和电信股恢复上涨的时间长度进行推断。

22. 尼尔森公司从欧洲、亚太、美洲和中东的47个市场对消费者进行调查，以确定影响他们购买食品地点的最主要的因素。用等级1（最低）到5（最高）表示。等级最高的因素是物有所值，平均分为4.32；等级次高的因素是更好地选择优质品牌和产品，平均分为3.78；等级最低的因素是使用可循环购物袋和包装，平均分为2.71（Nielsen website，2008年2月24日）。假设你受雇于一个食品连锁店进行类似的调查，以确定在北卡罗来纳州夏洛特市连锁店，消费者认为哪个因素是影响他们确定购买食品地点的最主要因素。

a. 你进行的这项研究的总体是什么？

b. 你如何搜集这项研究的数据？

24. 5名学生期中考试成绩的样本数据如下：72，65，82，90，76。下列表述中哪一个是正确的，哪一个因太空泛而受到怀疑？

a. 5名学生的平均期中成绩是77。

b. 参加考试的所有学生的平均期中成绩是77。

c. 参加考试的所有学生的平均期中成绩的估计值是77。

d. 一半以上参加考试的学生的成绩在70~85。

e. 如果这个样本中还包含其他5名学生，他们的成绩将在65~90。

第2章

描述统计学Ⅰ：表格法和图形法

CHAPTER 2

实践中的统计

高露洁 - 棕榄公司 ⊖
纽约州，纽约市

高露洁 - 棕榄（Colgate-Palmolive）公司于 1806 年在纽约开业时，还只是一家经营香皂和蜡烛的小商店。但今天，高露洁—棕榄公司已是一家遍布世界 200 多个国家和地区，拥有超过 40 000 名员工的大公司。除了最著名的品牌高露洁、棕榄、Ajax 和 Fab 外，同时还兼营 Mennen、希尔斯科学食品（Hill's Science Diet）和希尔斯处方食品（Hill's Prescription Diet）等产品。

高露洁 - 棕榄公司在其家用洗衣粉产品的生产过程质量保证计划中使用统计学方法。一个关键问题是顾客对盒装洗衣粉数量的满意度。相同尺寸的每一个盒子里都装入相同重量的洗衣粉，但是洗衣粉所占的体积受洗衣粉密度的影响。例如，如果洗衣粉的密度较大时，要达到所规定的包装重量，就只需要较少体积的洗衣粉。这样，当顾客打开包装盒时，盒子看上去显然没有装满。

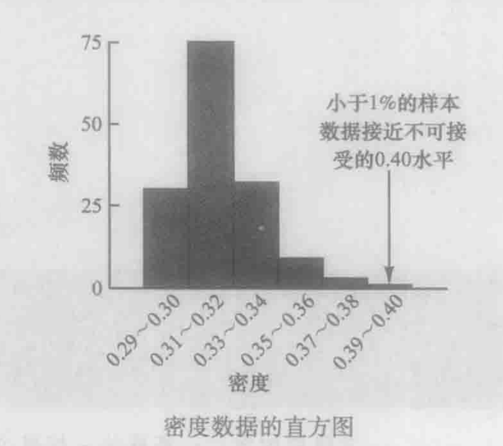
为了控制洗衣粉密度过大这个问题，需要对洗衣粉密度的可接受范围加以限制。定期抽取统计样本，并测量每一个洗衣粉样本的密度。然后将汇总数据提供给操作人员，以便在需要将密度保持在规定的质量规格内时，操作人员可以采取正确的措施。

根据在一周内采集的一个容量为 150 的样本，得到密度的频数分布和直方图如右图所示。密度水平超过 0.40 是不可以接受的上限。频数分布和直方图显示，所有产品的密度小于或等于 0.40 时，操作符合质量标准。检查这些统计汇总结果的管理人员对洗衣粉产品的质量感到满意。

在本章中，我们将学习有关描述统计的表格法和图形法，包括频数分布、条形图、直方图、茎叶显示、交叉分组表等内容。这些方法的目的是汇总数据，以便使统计数据易于理解和解释。

密度数据的频数分布

密度	频数
0.29 ~ 0.30	30
0.31 ~ 0.32	75
0.33 ~ 0.34	32
0.35 ~ 0.36	9
0.37 ~ 0.38	3
0.39 ~ 0.40	1
总计	150



密度数据的直方图

正如第 1 章所述，数据可以分为分类型或数量型，**分类型数据**（categorical data）是用标签或名称来识别项目的类型。**数量型数据**（quantitative data）是表示多少或大小的数值。本章介绍的是用于汇总分类型数据和数量型数据的表格法和图形法。人们在年度报告、报纸文章和研究报告中常可以看到数据的表格和图形汇总，因此，理解它们是怎样形成的以及如何解释它们至关重要。

我们首先讨论汇总单变量数据的表格法和图形法，接下来讨论汇总两个变量数据的表格法和图形法，它们在某种程度上揭示了两个变量之间的关系。术语**数据可视化**（data visualization）常常用于描述汇总和表述一个数据集信息的图形显示的效用。最后一节给出了数据可视化的简介和创建有效的图形显示的指导方针。

现代统计软件包提供了汇总数据和编制图形的强大功能，Minitab 和 Excel 是其中应用最广泛的两个软件包。在本章的附录中，我们将演示它们的部分功能。

⊖ 作者感谢为“实践中的统计”提供了案例的高露洁 - 棕榄公司质量保证部经理 William R. Fowle。

2.1 汇总分类变量的数据

2.1.1 频数分布

我们从频数分布（frequency distribution）的定义开始，讨论如何使用表格法和图形法汇总分类型数据。

频数分布

频数分布是一种数据的表格汇总，表示在几个互不重叠组别中的每一组项目的个数（即频数）。

我们用下面的例子来说明如何编制和解释分类型数据的频率分布。可口可乐、健怡可乐、胡椒博士、百事可乐和雪碧是5种受欢迎的软饮料。假设表2-1中的数据是50次购买软饮料的样本数据。

表2-1 50次软饮料购买的样本数据

可口可乐	可口可乐	可口可乐	雪碧	可口可乐
健怡可乐	胡椒博士	健怡可乐	胡椒博士	健怡可乐
百事可乐	雪碧	可口可乐	百事可乐	百事可乐
健怡可乐	可口可乐	雪碧	健怡可乐	百事可乐
可口可乐	健怡可乐	百事可乐	百事可乐	百事可乐
可口可乐	可口可乐	可口可乐	可口可乐	百事可乐
胡椒博士	可口可乐	可口可乐	可口可乐	可口可乐
健怡可乐	雪碧	可口可乐	可口可乐	胡椒博士
百事可乐	可口可乐	百事可乐	百事可乐	百事可乐
百事可乐	健怡可乐	可口可乐	胡椒博士	雪碧

为了编制这些数据的频数分布，我们计算表2-1中每一种软饮料的出现次数。可口可乐出现19次，健怡可乐出现8次，胡椒博士出现5次，百事可乐出现13次，雪碧出现5次。这些数字汇总在表2-2的频数分布中。

这个频数分布汇总说明了在50次购买软饮料的样本中，5种软饮料是如何分配的。它提供了比表2-1中的原始数据更多的内涵。观察这个频数分布，我们看到可口可乐是排在首位，百事可乐居第二位，第三位是健怡可乐，雪碧和胡椒博士并列第四。频数分布还揭示了这5种销售量最高的软饮料受欢迎程度的信息。

表2-2 购买软饮料的频数分布

软饮料	频数
可口可乐	19
健怡可乐	8
胡椒博士	5
百事可乐	13
雪碧	5
总计	50

2.1.2 相对频数分布和百分数频数分布

频数分布表明在几个互不重叠的组别中每一组项目的个数（即频数）。然而，我们往往对每一组的项目所占的比例或百分比更感兴趣。一组的相对频数是属于该组别的项目个数占总数的比例。对一个有n个观测值的数据集，每一组的相对频数由下式给出。

相对频数

$$\text{组的相对频数} = \frac{\text{组的频数}}{n} \tag{2-1}$$

一组的百分数频数是相对频数乘以100。

相对频数分布（relative frequency distribution）是每一组相对频数数据的表格汇总。**百分数频数分布**（percent frequency distribution）是每一组百分数频数数据的表格汇总。表2-3是软饮料数据的相对频数分布的百分数频数分布。在表2-3中，我们看到可口可乐的相对频数为19/50 = 0.38，健怡可乐的相对频数为8/50 = 0.16，依此类推。从百分数频数分布中，我们看到在购买的软饮料中，有38%是可口可乐，有16%是健怡可乐，等等。我们还注意到前三类软饮料占购买量的80%（=38% +26% +16%）。

表2-3 购买软饮料的相对频数和百分数频数分布

软饮料	相对频数	百分数频数
可口可乐	0.38	38
健怡可乐	0.16	16
胡椒博士	0.10	10
百事可乐	0.26	26
雪碧	0.10	10
总计	1.00	100

2.1.3 条形图和饼形图

条形图（bar chart）[⊖]是一种图形方法，用来描绘已汇总的分类型数据的频数分布、相对频数分布或百分数频数分布。在图形的轴上（通常是横轴），我们规定了用来对数据分组（类别）的标记，在图形的另一轴上（通常是纵轴），标出频数、相对频数或百分数频数的刻度。然后，将用一个固定宽度的长条绘制在每一组的标记上，将这个长条的高度延伸，直到达到该组的频数、相对频数或百分数频数。对于分类型数据，应将这些长条分隔开，以强调每一组是相互独立的事实。图2-1是50次购买软饮料情况的频数分布的条形图。注意图形是如何显示出可口可乐、百事可乐和健怡可乐成为最受欢迎的品牌的。

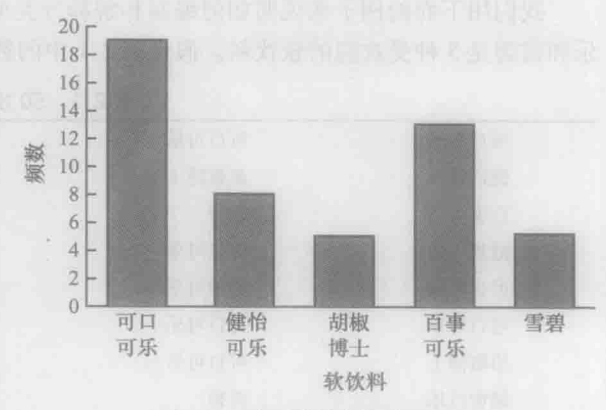


图2-1 购买软饮料的条形图

饼形图（pie chart）是另一种描绘分类型数据的相对频数和百分数频数分布的图形方法。为了绘制饼形图，我们首先画一个圆来代表所有的数据，然后用相对频数把圆细分成若干扇形部分，这些扇形与每一组的相对频数相对应。例如，因为一个圆有360°，而可口可乐的相对频数是0.38，则饼形图中代表可口可乐的扇形部分为0.38×360°=136.8°，而代表健怡可乐的扇形部分为0.16×360°=57.6°，对其他组进行相似的计算，就得到图2-2中的饼形图。显示每一个扇形部分的数值可以是频数、相对频数或百分数频数。

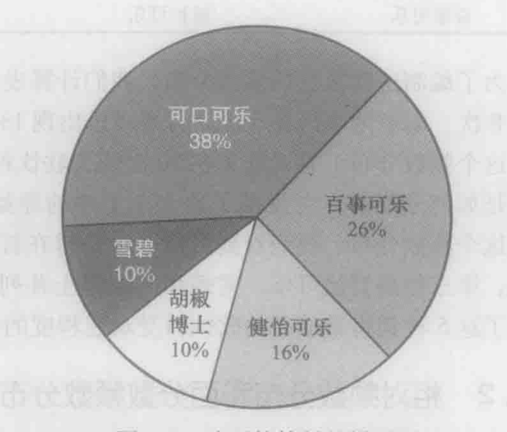


图2-2 购买软饮料的饼形图

包含颜色、阴影、图例、文本字体和三维透视图使用的许多选项可以增加条形图和饼形图的视觉显示。当仔细使用时，这些选项可以提供更有效的显示。但是，这并不总是实情。例如，考虑图2-3中的软饮料的三维饼形图。将其与图2-2中较为简单的显示相比，三维透视图没有增加新的理解。事实上，由于你不得不以一个角度观看图2-3中的三维饼形图，而不是直视，因此，形象化可能较为困难。图2-3图例的使用也使你的眼睛在键盘和图之间来回地移动。图2-2中较为简单的图在饼形图上直接显示百分比和类别，更有效。

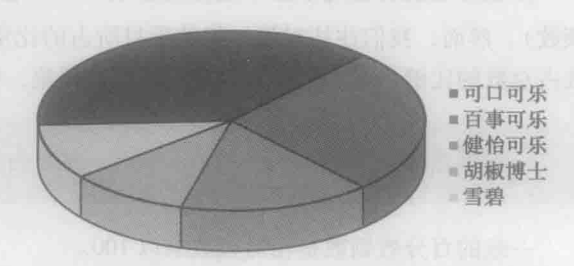


图2-3 购买软饮料的三维饼形图

一般地，饼形图不是呈现对比的百分比的最佳途径。研

⊖ 在质量管理应用中，用条形图来鉴明问题最重要的原因。当条形图按高度降序顺序从左到右依次排列时，最频繁发生的原因出现在第一位时，这种条形图称作帕雷托图。这种图形是以创立者意大利经济学家维尔弗雷多·帕雷托（Vilfredo Pareto）的名字命名的。

究表明人们更愿意准确判断长度间的差异，而非角度（或部分）间的差异。当进行这样的比较时，我们建议你使用类似图2-1的条形图。在2.5节我们给出了创建有效的视觉显示的更多指导。

注释和评论

1. 在一个频数分布中，组数通常与在数据中发现的类别数相同，就像本节中购买软饮料数据中的情况一样。数据只涉及5种软饮料，每一种软饮料定义为一个独立的频数分布组别。涉及所有的软饮料的数据将需要很多类别，其中的大部分组别中只有很少的购买次数。大多数统计专家建议把频数较小的组合并到一个叫"其他"的综合组中去。只有5%或更少频数的组通常被这样处理。
2. 在任何频数分布中，频数的总和总是等于观测值的数目。在任何相对频数中，相对频数的总和总是等于1.00，在任何百分数频数分布中，百分数的总和总是等于100。

练习

方法

2. 下面给出了一部分相对频数分布：

组别	相对频数
A	0.22
B	0.18
C	0.40
D	

a. 组别D的相对频数是多少？

b. 若样本容量为200，则组别D的频数是多少？

c. 求频数分布。

d. 求百分数频数分布。

应用

4. 2010～2011年电视季，收视率最高的5个辛迪加电视节目是：《命运之轮》（*Wheel of Fortune*，WoF）、《好汉两个半》（*Two and Half Men*，THM）、《危险边缘》（*Jeopardy*，Jep）、《朱迪法官》（*Judge Judy*，JJ）和《奥普拉脱口秀》（*Oprah Winfrey Show*，OWS）（Nielsen Media Research website，2012年4月16日）。由50名电视观众组成一个样本，表明首选节目的数据如下。

WoF	Jep	JJ	Jep	THM
THM	WoF	OWS	Jep	THM
Jep	OWS	WoF	WoF	WoF
WoF	THM	OWS	THM	WoF
THM	JJ	JJ	Jep	THM
OWS	OWS	JJ	JJ	Jep
JJ	WoF	THM	WoF	WoF
THM	THM	WoF	JJ	JJ
Jep	THM	WoF	Jep	Jep
WoF	THM	OWS	OWS	Jep

a. 这些数据是分类型的还是数量型的？

b. 编制这些数据的频数分布和百分数频数分布。

c. 绘制这些数据的条形图和饼形图。

d. 以样本为依据，哪个电视节目拥有最多的电视观众？哪个位居第二？

6. 尼尔森媒体研究提供了电视史上收视率前25名的单个节目的清单（《世界年鉴》，2012年）。下面是制作收视率前25名电视节目的电视网的数据。

CBS	CBS	NBC	FOX	CBS
CBS	NBC	NBC	NBC	ABC
ABC	NBC	ABC	ABC	NBC
CBS	NBC	CBS	ABC	NBC
NBC	CBS	CBS	ABC	CBS

a. 编制数据的频数分布、百分数频数分布以及条形图。

b. 在提供收视率最高的电视节目方面，哪个电视网做得最好？比较ABC，CBS和NBC的表现。

8. 美国棒球名人堂位于美国纽约州的古柏镇（Cooper-town）。由入选名人堂的55名球员组成一个样本，样本数据如下所示。每一个观测值表示入选名人堂的球员的主要位置：投手（P），接手（H），一垒手（1），二垒手（2），三垒手（3），游击手（S），左外场手（L），中外场手（C）和右外场手（R）。

L	P	C	H	2	P	R	1	S	S	1	L	P	R	P
P	P	P	R	C	S	L	R	P	C	C	P	P	R	P
2	3	P	H	L	P	1	C	P	P	P	S	1	L	R
R	1	2	H	S	3	H	2	L	P					

a. 用频数分布和相对频数分布来汇总数据。

b. 哪个位置的名人堂球员最多？

c. 哪个位置的名人堂球员最少？

d. 哪个外场位置（L、C或R）的名人堂球员最多？

e. 比较内场人数（1、2、3和S）与外场人数（L、C和R）。

10.《金融时报》/哈里斯民意调查对6个欧洲国家和美国的成年人每月进行一次民意在线调查。有1 015名成年人参加了2008年1月份的调查，其中一个问题是："您如何评价联邦银行在处理金融市场信用问题时的表现？"可能的答案有：优秀、良好、中等、较差和极差（Harris Interactive website，2008年1月）。这1 015人的回答数据保存在名为FedBank的文件中。

a. 编制频数分布。

b. 编制百分数频数分布。

c. 绘制百分数频数分布的条形图。

d. 美国成年人如何评价联邦银行处理金融市场中的信用问题。

e. 在西班牙，1 114名成年人被问及"您如何评价对欧洲中央银行在处理金融市场信用问题的表现？"得到如下的百分数频数分布：

等级	百分数频数分布
优秀	0
良好	4
中等	46
较差	40
极差	10

比较西班牙和美国的结果。

2.2 数量型数据汇总

2.2.1 频数分布

正如第2.1节中定义的那样，频数分布是表示在几个互不重叠的组中每一组所包含的项目个数（或频数）的表格汇总。这个定义也适用于数量型数据。然而，对于数量型数据，在定义用于频数分布的互不重叠的组时，我们必须更加小心。

例如，考虑表2-4中的数量型数据。这些样本数据是一家名叫Sanderson—Clifford的小型会计师事务所对20位客户完成年末审计所需的时间（单位：天）。对于数量型数据，确定频数分布的组时，三个必要的步骤是：

表2-4 年末审计时间 （单位：天）

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

（1）确定互不重叠组的组数。

（2）确定每组的宽度。

（3）确定组限。

让我们通过编制表2-4的审计时间数据的频数分布来演示这些步骤。

组数 组是通过对数据规定范围而形成的，这个规定的范围用于对数据进行分组。作为一般性原则，我们建议使用5~20个组。如果数据项较少，只用5组或6组就可以汇总数据。如果数据项较多，通常需要较多的组。分组的目的是用足够多的组来显示数据的变异性，而不是有这么多的组却只包含很少的数据项。因为表2-4中的数据项相对较少（$n=20$），我们决定选择编制5组的频数分布。

组宽 编制数量型数据频数分布的第二步是选择组宽。作为一般性原则，我们建议每组的宽度相同[⊖]。因此，对组数和组宽的选择就不能独立决定。较大的组数意味着较小的组宽，反之亦然。为了确定一个近似的组宽，我们从确定数据的最大值和最小值开始。一旦确定了期望的组数，我们就可以用下面的表达式来确定近似的组宽。

$$\text{近似组宽} = \frac{\text{数据最大值} - \text{数据最小值}}{\text{组数}} \tag{2-2}$$

由式（2-2）给出的近似组宽，可以根据编制频数分布的人的偏好取整为更方便的值。例如，近似组宽为9.28，可以简单地取整为10，因为10作为组宽在编制频数分布时更方便。

对于年末审计时间数据，最大值是33，最小值是12。因为我们决定用5组来汇总数据，由式（2-2）可计算出近似组宽是（33-12）/5=4.2。因此，我们决定以5天作为频数分布的组宽。

⊖ 每组的宽度相同，可减少使用者的不恰当诠释机会。

在实践中，组数和近似组宽要通过反复试验确定。一旦选定了一个可能的组数，根据式（2-2）就可以找出近似组宽。这个过程能对不同的组数反复进行。最终，分析人员利用判断来确定组数和组宽的组合，得到最佳的汇总数据的频数分布。⊖

对于表2-4中的审计时间数据，在决定了使用5组，每一组以5天为组宽后，下一步的工作就是规定每一组的组限。

组限　选择组限必须使每一个数据值属于且只属于一组。下组限定义为被分到该组的最小可能的数据值，上组限定义为被分到该组的最大可能的数据值。在编制分类型数据的频数分布时，我们不需要规定组限，因为每一数据项都会自然地落入分隔开的组内。但对于数量型数据，比如表2-4中的审计时间，组限对确定每个数据项的归属很有必要。

对于表2-4中的审计时间数据，我们对第一组选择10天为下组限和14天为上组限。该组在表2-5中记为10～14。最小数据值12包含在10～14组。然后，我们对下一组选择15天为下组限和19天为上组限。我们继续确定下、上组限，直到获得全部的5个组：10～14，15～19，20～24，25～29和30～34。最大数据值33包含在30～34组。相邻两组的下组限之差就是组宽。利用前两个下组限10和15，我们看到，组宽为15－10＝5。

一旦确定了组数、组宽和组限，通过统计属于每一组的数据值的个数，我们可以得到频数分布。例如，表2-4中的数据显示，有4个值（12，14，14和13）属于10～14组。因此，10～14组的频数是4。对15～19，20～24，25～29和30～34各组继续计数过程，得到表2-5中的频数分布。利用这个频数分布，我们可以观察到：

（1）最频繁发生的审计时间在15～19天这一组，在20个审计时间中有8个属于这一组。

（2）只有一次审计需要30天或更多的时间。

根据人们对频数分布的兴趣，可能得出其他结论。一个频数分布的价值就在于它提供了对数据的深刻理解，而直接观察原始的未经组织的数据，是无法获得这种深刻理解的。

组中值　在一些应用中，我们想知道数量型数据频数分布的组中值，**组中值**（class midpoint）是下组限和上组限的中间值。对审计时间数据，5个组的组中值分别是12，17，22，27和32。

表2-5　审计时间数据的频数分布

审计时间（天）	频数
10～14	4
15～19	8
20～24	5
25～29	2
30～34	1
总计	20

2.2.2　相对频数分布和百分数频数分布

我们以与分类型数据同样的方式来定义数量型数据的相对频数和百分数频数分布。首先，我们记得，相对频数是属于一个组的观察值的比例。对于n个观察值

$$组的相对频数=\frac{组频数}{n}$$

组的百分数频数是相对频数乘以100。

根据表2-5中的组频数和$n=20$，得到表2-6中的审计时间数据的相对频数分布和百分数频数分布。注意，有0.40或40%的审计需要15～19天时间，只有0.05或5%的审计需要30天或更多时间。通过使用表2-6，我们还能够得到其他的解释和理解。

表2-6　审计时间数据的相对和百分数频数分布

审计时间（天）	相对频数	百分数频数
10～14	0.20	20
15～19	0.40	40
20～24	0.25	25
25～29	0.10	10
30～34	0.05	5
总计	1.00	100

2.2.3　打点图

打点图（dot plot）是一种最简单的数据图形汇总方式。横轴是数据的值域，每一个数据值由位于横轴上的点表示。图2-4是表2-4中审计时间数据的打点图，3个点位于横轴刻度18之上，表明有3

⊖　对一个数据集，不存在最佳的频数分布。不同的人可以编制不同的，但同样可接受的频数分布。频数分布的目的是揭示自然的分组和数据的变异性。

次审计时间为18天。打点图能够显示数据的细节，且有利于比较两个或更多变量的数据分布。

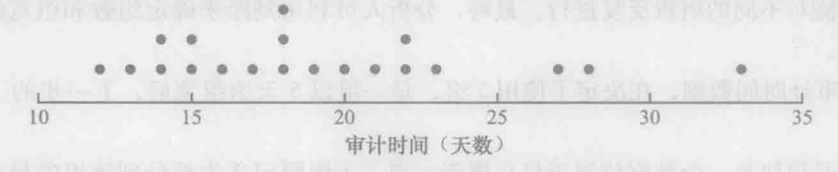


图 2-4 审计时间数据的打点图

2.2.4 直方图

直方图（histogram）是一种常用的数量型数据的图形描述方式。由先前已汇总出的频数分布、相对频数分布或百分数频数分布等资料可绘制直方图。将所关心的变量放置在横轴上，将频数、相对频数或百分数频数放置在纵轴上，就可以绘制一个直方图。每组的频数、相对频数或百分数频数用一个长方形绘制，长方形的底放置在横轴上，以组宽为底，以每组相应的频数、相对频数或百分数频数为高。

图2-5是审计时间数据的直方图。我们注意到，最大频数的组由15～19天这一组的长方形表示，长方形的高度表示这一组的频数是8。这些数据的相对频数分布或百分数频数分布的直方图看起来与图2-5中的直方图一样，只是纵轴用相对频数或百分数频数标记。

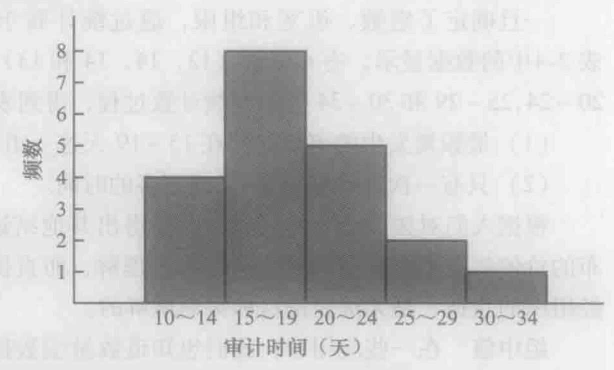


图 2-5 审计时间数据的直方图

正如图2-5所示，直方图中邻近的长方形是互相连接的。与条形图不同，直方图相邻组的长方形之间没有自然的间隔。这种形式是直方图的惯例。因为审计时间数据的各组表示为10～14，15～19，20～24，25～29和30～34，各组间从14～15，19～20，24～25和29～30有一个单位的间隔，在绘制直方图时可以消除该间隔。在审计时间数据的直方图中消除了组间的间隔，有助于说明即使数据是近似的，在第一组的下限和最后一组的上限间所有的值都是可能的。

直方图的一个最重要应用是提供了分布形态的信息。图2-6是四个根据相对频数分布绘制的直方图，图2-6a显示一个数据集的直方图有一定程度的左偏。这个直方图说明，如果图形的尾部向左延伸一些，则图形左偏。考试成绩是这种直方图的典型应用。因为没有成绩在100%之上，大多数成绩又常常在70%之上，只有极少数的成绩很低。图2-6b显示一个数据集的直方图有一定程度的右偏。这个直方图说明，如果图形的尾部向右延伸一些，则图形右偏。像房屋价格的数据可能是这种直方图的例子，少数昂贵的住宅造成右尾偏斜。

图2-6c是一个对称的直方图。在对称的直方图中，左尾和右尾的形状相同。在应用中得到数据的直方图永远不会完全对称，但许多应用中的直方图可能大致对称。SAT分数、人的身高和体重等数据得到的直方图是大致对称的。图2-6d显示直方图严重右偏。这个直方图是根据一家妇女服饰店一整天的销售量数据绘制出来的。在商务与经济应用中得到的数据，常常使直方图右偏。例如，房屋价格、工资、销售量等数据，常常导致直方图右偏。

2.2.5 累积分布

对频数分布略加变化，可得到数值型数据的另一种表格汇总方式——**累积频数分布**（cumulative frequency distribution）。累积频数分布使用编制频率分布的组数、组宽和组限。但是，累积频数分布表示的是小于或等于每一组上组限的数据项个数，而不是表示每一组的频数。表2-7前两列给出了审计时间数据的累积频数分布。

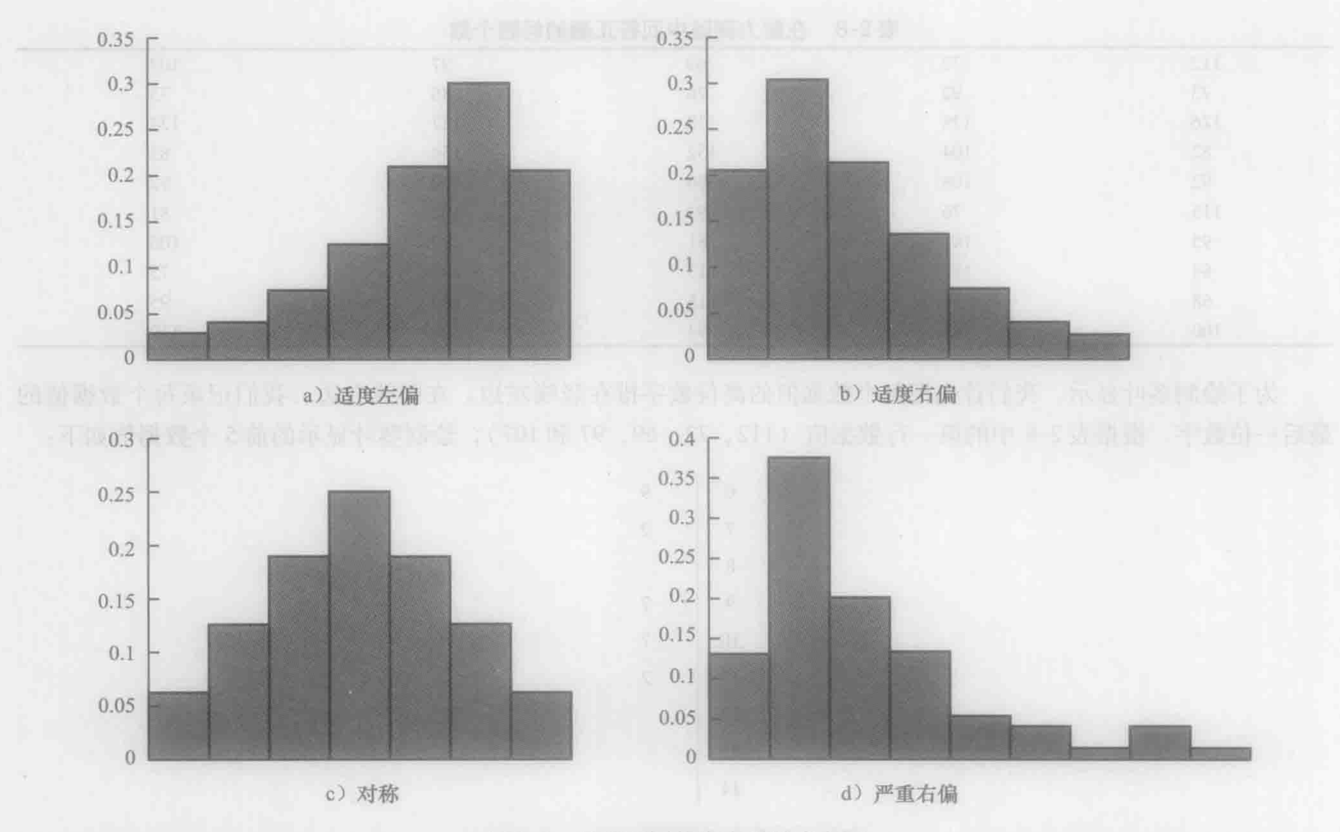


图2-6　呈现不同偏度水平的直方图

表2-7　审计时间数据的累积频数、累积相对频数和累积百分数频数分布

审计时间（天）	累积频数	累积相对频数	累积百分数频数
小于或等于14	4	0.20	20
小于或等于19	12	0.60	60
小于或等于24	17	0.85	85
小于或等于29	19	0.95	95
小于或等于34	20	1.00	100

为了解释累积频数是如何确定的，考虑“小于或等于24”的这一组。这一组的累积频数只是数据中小于或等于24的所有组的频数之和。对表2-5中的频数分布而言，组10～14，15～19和20～24的频数之和为4+8+5=17，它表明有17个数据小于或等于24。因此，这一组的累积频数是17。另外，表2-7中的累积频数分布表明，有4次审计在14天内完成，有19次审计在29天内完成。

最后，我们注意到，**累积相对频数分布**（cumulative relative frequency distribution）和**累积百分数频数分布**（cumulative percent frequency distribution）分别表示数据值小于或等于每一组上组限的数据项的比例和百分数。我们既可以对相对频数分布中的相对频数求和，也可以用累积频数除以数据项总数，来计算累积相对频数分布。用后一种方法，我们将表2-7中第2列的累积频数除以数据项总数（$n=20$），可得到第3列的累积相对频数分布。累积百分数频数由累积相对频数乘以100得出。累积相对频数和累积百分数频数分布显示，有0.85或85%的审计在24天内完成，有0.95或95%的审计在29天内完成，等等。

2.2.6　茎叶显示

茎叶显示（stem-and-leaf display）是同时用于显示数据的等级顺序和分布形状的图形显示。为了说明茎叶显示的使用，考虑表2-8中的数据。这些数据是Haskens制造公司最近对50名应聘某职位的人提出150个能力测验问题进行面试时，回答正确的问题个数。

表 2-8 在能力测验中回答正确的问题个数

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

为了绘制茎叶显示，我们首先把每个数据值的高位数字排在竖线左边。在竖线右边，我们记录每个数据值的最后一位数字。根据表 2-8 中的第一行数据值（112，72，69，97 和 107），绘制茎叶显示的前 5 个数据值如下：

6 | 9
7 | 2
8 |
9 | 7
10 | 7
11 | 2
12 |
13 |
14 |

例如，数值 112 的最高数字 11 在竖线左边，最后一位数字 2 在竖线右边。类似地，数值 72 的最高数字 7 在竖线左边，最后一位数字 2 在竖线右边。对应每个数据值的最高数字，继续把它的最后一位数字放置在竖线的右边，得到如下结果：

6 | 9 8
7 | 2 3 6 3 6 5
8 | 6 2 3 1 1 0 4 5
9 | 7 2 2 6 2 1 5 8 8 5 4
10 | 7 4 8 0 2 6 6 0 6
11 | 2 8 5 9 3 5 9
12 | 6 8 7 4
13 | 2 4
14 | 1

这样组织数据后，对每一行的数字排出等级顺序就简单了。排序后的茎叶显示如下：

6 | 8 9
7 | 2 3 3 5 6 6
8 | 0 1 1 2 3 4 5 6
9 | 1 2 2 2 4 5 5 6 7 8 8
10 | 0 0 2 4 6 6 6 7 8
11 | 2 3 5 5 8 9 9
12 | 4 6 7 8
13 | 2 4
14 | 1

竖线左边的数字（6，7，8，9，10，11，12，13和14）是茎，竖线右边的每个数字就是一个叶。例如，考虑第一行，6是茎的值，8和9是两个叶的值：

```
6 | 8  9
```

它表示有两个数据值的首位数字是6，叶表示数据值是68和69。类似地，第二行是

```
7 | 2  3  3  5  6  6
```

它表示有6个数据值其首位数字是7，叶表示数据值是72，73，73，75，76和76。

根据茎叶显示，我们关注所显示出的形状。让我们用一个长方形围住每个茎的叶。这样做，我们得到下图：

```
 6 | 8  9
 7 | 2  3  3  5  6  6
 8 | 0  1  1  2  3  4  5  6
 9 | 1  2  2  2  4  5  5  6  7  8  8
10 | 0  0  2  4  6  6  6  7  8
11 | 2  3  5  5  8  9  9
12 | 4  6  7  8
13 | 2  4
14 | 1
```

将上图逆时针方向旋转90°，所得到的图形与按60～69，70～79，80～89等分组的直方图非常相似。

尽管茎叶显示可能与直方图提供相同的信息，但它还有以下两个主要优点：

（1）茎叶显示易于用手绘制。

（2）在一个组内，由于茎叶显示提供了实际的数据值，因此茎叶显示比直方图提供更多的信息。

就像频数分布或直方图没有绝对的组数一样，茎叶显示也没有绝对的行数或茎数。如果我们认为初始的茎叶显示聚集了太多的数据，我们能够把每个高位数字拆分为两个或更多的茎，从而很容易扩展这个显示。例如，对每一高位数字使用两个茎，我们把所有以0，1，2，3和4结尾的数据放在一行，而把所有以5，6，7，8和9结尾的数据放在另一行。下列扩展的茎叶显示[⊖]演示了这种方法。

```
 6 | 8  9
 7 | 2  3  3
 7 | 5  6  6
 8 | 0  1  1  2  3  4
 8 | 5  6
 9 | 1  2  2  2  4
 9 | 5  5  6  7  8  8
10 | 0  0  2  4
10 | 6  6  6  7  8
11 | 2  3
11 | 5  5  8  9  9
12 | 4
12 | 6  7  8
13 | 2  4
13 |
14 | 1
```

⊖ 在扩展的茎叶显示中，每当茎值被重复两次时，则第一个茎值对应于叶值0～4，第二个茎值对应于叶值5～9。

注意，数值72，73和73的叶值在0~4范围内，并且用7的第一个茎值来显示。数值75，76和76的叶值在5~9范围内，用7的第二个茎值来显示。这个扩展的茎叶显示与间隔为65~69，70~74，75~79等的频数分布很相似。

前面的例子说明了三位数数据的茎叶显示。对于超过三位数的数据，也能作出茎叶显示。例如，考虑下列数据，它们是一个快餐店15个周的汉堡包的销售数量。

1 565	1 852	1 644	1 766	1 888	1 912	2 044	1 812
1 790	1 679	2 008	1 852	1 967	1 954	1 733	

这些数据的茎叶显示如下：

叶单位=10

15	6			
16	4	7		
17	3	6	9	
18	1	5	5	8
19	1	5	6	
20	0	4		

注意，这里用一个单一数字来表示每一个叶值，并且每个数据值只用前三位数字来绘制茎叶显示。在茎叶显示的顶部，我们规定了叶单位为10。为了说明如何解释茎叶显示中的数值，考虑第一个茎值15，以及它的叶值6。把它们连起来，我们得到数156。要得到原始观察值的一个近似值，我们需要把它乘以叶单位的值10。因此，156×10=1 560就是用来绘制茎叶显示的原始数据值的近似值。虽然从茎叶显示中得出精确数据值是不可能的，但是根据每个叶值使用一位数表示的惯例[⊖]，可以保证我们对位数较多的数据也能绘制出茎叶显示。对于没有给出叶单位的茎叶显示，我们就假设它等于1。

注释和评论

1. 条形图和直方图本质上是同一事物，它们都是频数分布数据的图形表示。直方图是各纵条之间没有间隔的条形图。有些离散的数量型数据，各纵条之间有间隔是合适的。例如，考虑大学中注册学生的班级数，这些数据只能取整数，像1.5和2.73等的中间值是不可能存在的。对于连续的数量型数据，如表2-4的审计时间，各纵条之间的间隔是不合适的。
2. 对于数量型数据，适当的组限依赖于数据的精度水平。例如，对表2-4中的审计时间数据，我们取整数值为组限。如果数据近似到最接近的1/10天（如12.3，14.4等），那么组限将以1/10天表示。例如，第一组将会设为10.0~14.9。如果数据近似到最接近的1/100天（如12.34，14.45等），那么组限将以1/100天表示。例如，第一组将会设为10.00~14.99。
3. 开口组是指只有一个下组限或上组限的组。例如，在表2-4中的审计时间数据中，假设有两次审计时间分别为58天和65天，我们将不再继续以5为组宽，将组延伸到35~39，40~44，45~49等，而是以一个开口组"35或35以上"来简化频数分布，这一组的频数为2。开口组经常出现在分布的最上端，有时开口组也出现在分布的最下端，偶尔也在两端出现。
4. 累积频数分布的最后一个数据项总等于观测值的总数。累积相对频数分布的最后一个数据项总等于1.00，累积百分数频数分布的最后一个数据项总等于100。

⊖ 在茎叶显示中使用一个单一数字来表示每个叶。叶单位是为了得到原始数据的近似值，而在茎叶显示中应该乘以的数值。叶单位可以是100，10，1，0.1等。

练　习

方法

12. 考虑下面的频数分布：

组	频数	组	频数
10～19	10	40～49	7
20～29	14	50～59	2
30～39	17		

编制累积频数分布和累积相对频数分布。

14. 考虑下列数据：

8.9	10.2	11.5	7.8	10.0	12.2	13.5	14.1	10.0	12.2
6.8	9.5	11.5	11.2	14.9	7.5	10.0	6.0	15.8	11.5

a. 绘制打点图。

b. 编制频数分布。

c. 编制百分数频数分布。

16. 绘制下列数据的茎叶显示，叶单位为10。

1 161	1 206	1 478	1 300	1 604	1 725	1 361	1 422
1 221	1 378	1 623	1 426	1 557	1 730	1 706	1 689

应用

18. 投考者的不足需要学区支付较高的薪水和额外的补助以吸引和留住学区的负责人。纽约州大罗切斯特地区20个学区负责人的基本年薪数据（单位：1 000美元）如下所示（The Rochester Democrat and Chronicle，2008年2月10日）。

187	184	174	185
175	172	202	197
165	208	215	164
162	172	182	156
172	175	170	183

下面按150～159，160～169等分组。

a. 编制频数分布。

b. 编制百分数频数分布。

c. 编制累积百分数频数分布。

d. 绘制基本年薪的直方图。

e. 数据是否呈现偏斜？请解释。

f. 基本年薪超过200 000美元的负责人占有多大比例？

20. 伦敦经济学院和哈佛商学院进行了一项关于首席执行官如何度过他们的一天的研究。研究发现首席执行官每周在会议上大约平均花18个小时，其中不包括电话会议、商务会餐和公众活动（《华尔街日报》，2012年2月14日）。下表是25位首席执行官每周花在会议上的时间的样本数据（单位：小时）。

14	15	18	23	15
1	20	13	15	23
23	21	15	20	21
16	15	18	18	19
19	22	23	21	12

a. 每周花在会议上的时间的最小值和最大值分别是多少？

b. 以2小时为组宽，编制这些数据的频数分布和百分数频数分布。

c. 绘制直方图，并评价分布的形态。

22.《企业家》杂志利用诸如增长率、场所数量、启动资金和财政稳定等工作指标对特许加盟商评定等级。美国特许加盟商20强的场所数量如下表所示。

特许加盟商	美国的场所数量
Hampton Inns	1 864
ampm	2 183
麦当劳	32 805
7-Eleven 公司	37 496
Supercuts	2 13
Days Inn	1 877
Vanguard Cleaning System	2 155
Serpro	1 522
Subway	34 871
Denny's 公司	1 668
Jan-Pro Franchising Intl 公司	12 394
Hardee's	1 901
Pizza Hut 公司	13 281
Kumon Math & Reading Centers	25 199
Dunkin's Donuts	9 947
肯德基公司	16 224
Jazzercise 公司	7 683
Anytime Fitness	1 618
Matco Tools	1 431
Stratus Building Solutions	5 018

下面按0～4 999，5 000～9 999，10 000～14 999及以上分组，并回答下列问题。

a. 编制这些顶级特许加盟商美国的场所数量的频数分布和百分数频数分布。

b. 绘制这些数据的直方图。

c. 评价分布的形态。

24.《财经杂志》列出了工作的最佳职业机会是令人愉快、薪金满意以及迄今为止在此工作10年之久（《财经杂志》，2009年11月）。下面是拥有2～7年工作经验的20个最佳职业机会，其员工薪金的

中位数收入和最高收入数据（单位：1 000 美元）。

职业	中位数收入	最高收入
业务经理	81	157
注册会计师	74	138
计算机安全顾问	100	138
通讯主管	78	135
财务分析员	80	109
财务主管	121	214
财务研究分析员	66	155
酒店总经理	77	146
人力资源经理	72	111
投资银行业务员	106	221
IT 商业分析员	83	110
IT 项目经理	99	140
市场经理	77	126
质量保证经理	80	122
商品经销代理	67	125
高级内部审计师	76	106
软件开发员	79	116
软件项目经理	110	152
系统工程师	87	130
文档工程师	67	100

对中位数收入和最高收入编制茎叶显示，评价这些职业的收入状况。

26. 2011 年辛辛那提 Flying Pig 半程马拉松（13.1 英里）有 10 897 名参赛者完成了比赛（Cincinnati Flying Pig Marathon website）。40 名半程马拉松参赛者的年龄数据如下表所示。

49	33	40	37	56
44	46	57	55	32
50	52	43	64	40
46	24	30	37	43
31	43	50	36	61
27	44	35	31	43
52	43	66	31	50
72	26	59	21	47

a. 绘制扩展的茎叶显示。

b. 哪个年龄组的参赛者最多？

c. 哪个年龄出现的次数最多？

2.3 用表格方式汇总两个变量的数据

本章目前为止，我们已集中讨论了利用表格法和图形法对一个分类或数量变量的数据进行汇总。管理人员和决策者往往需要汇总两个变量的数据，以揭示变量间关系。本节，我们介绍如何编制两个变量数据的表格汇总。

2.3.1 交叉分组表

交叉分组表（crosstabulation）是一种汇总两个变量数据的方法。虽然两个变量可以是分类的或数量的，但一个变量是分类的而另一个变量是数量的交叉分组表最为常见。通过考虑根据下面《Zagat 饭店评论》（*Zagat's Restaurant Review*）数据的应用，我们来说明后一种情形的交叉分组表。由位于洛杉矶地区的 300 家饭店组成一个样本，搜集它们的质量等级和代表性餐价数据。表 2-9 给出了前 10 家饭店的数据。质量等级是一个分类变量，等级类别有好、很好和优秀；餐价是一个数量变量，变化范围从 10～49 美元。

这一应用的数据交叉分组表如表 2-10 所示。左边栏和顶部边栏的标记确定了两个变量的组别。左边栏的标记（好、很好和优秀）对应着质量等级变量的三个组。顶部边栏的列标记（10～19 美元，20～29 美元，30～39 美元和 40～49 美元）对应着餐价变量的四个组。[⊖]样本中的每个饭店都给出了质量等级和餐价。因此，样本中的每个饭店都与交叉分组表中的某一列和某一行的交叉单元相联系。例如，饭店 5 被认定为质量等级为很好，33 美元的餐价。它属于表 2-10 中第 2 行和第 3 列的单元。在编制交叉分组表时，我们只需简单地计算出属于交叉分组表每个单元的饭店数。

表 2-9 300 家洛杉矶饭店的质量等级和餐价

饭店	质量等级	餐价（美元）
1	好	18
2	很好	22
3	好	28
4	优秀	38
5	很好	33
6	好	28
7	很好	19
8	很好	11
9	很好	23
10	好	13
⋮	⋮	⋮

虽然在表 2-10 中，餐价变量使用四个组来编制交叉分组表，但是对于餐价变量可以使用更少或更多的组，来构建质量等级和餐价

⊖ 在编制交叉分组表时，对数量变量进行分组可使我们将数量变量看成一个分类变量。

的交叉分组表。涉及在交叉分组表中如何对数量变量的数据进行分组的问题，与利用数量变量来编制频数分布时确定组数的问题相类似。对于这个应用，餐价变量的四个组被认为是揭示质量等级和餐价之间关系的合理组数。

观察表2-10，我们看到，样本中质量等级为很好且餐价在20～29美元的饭店最多（64）。质量等级为优秀且餐价在10～19美元的饭店只有两家。其他频数也可以进行类似的解释。另外，我们注意到交叉分组表的右边和最后一行分别给出了质量等级和餐价的频数分布。从右边的频数分布中我们看到，质量等级为好的饭店有84家，很好的饭店有150家和优秀的饭店66家。类似地，最后一行是餐价变量的频数分布。

表2-10　300家洛杉矶饭店的质量等级和餐价数据的交叉分组表

质量等级	餐价（美元）				
	10～19	20～29	30～39	40～49	总计
好	42	40	2	0	84
很好	34	64	46	6	150
优秀	2	14	28	22	66
总计	78	118	76	28	300

用交叉分组表右边栏的行总计数除以饭店总数300，可以得到质量等级变量的相对频数和百分数频数分布。

质量等级	相对频数	百分数频数
好	0.28	28
很好	0.50	50
优秀	0.22	22
总计	1.00	100

从百分数频数分布，我们看到有28%的饭店的质量等级是好，有50%的质量等级是很好，有22%的质量等级是优秀。

用交叉分组表最后一行的列总计数除以饭店的总数300，可以得到餐价变量的相对频数和百分数频数分布。

餐价（美元）	相对频数	百分数频数
10～19	0.26	26
20～29	0.39	39
30～39	0.25	25
40～49	0.09	9
总计	1.00	100

注意，由于数值经过四舍五入后求和，因此列的相对频数值之和不一定精确地等于1.00，而且百分数频数分布值之和不一定精确等于100。从百分数频数分布我们看到，有26%的饭店餐价在最低价格组（10～19美元），有39%饭店的餐价在次低价格组，等等。

从交叉分组表边栏得到的频数分布和相对频数分布，可以提供每一个变量单独的信息，但它们不能提供变量间关系的任何信息。交叉分组表的主要价值在于提供了变量间关系的深刻含义。再次查看表2-10的交叉分组表，揭示出较高的餐价与较高的质量等级相联系，而较低的餐价对应于较低的质量等级。

把交叉分组表中的项目转换成行百分数或列百分数，对变量间关系提供了更深入的了解。对行百分数，表2-10中的每个频数除以对应的行总计数，所得的结果显示在表2-11中。表2-11中的每一行是同一质量等级的餐价的百分数频数分布。对于质量等级最低（好）的饭店，我们看到最大的百分数是较便宜的饭店（50%的饭店餐价是10～19美元，47.6%的饭店餐价是20～29美元）。对于质量等级最高（优秀）的饭店，我们看到最大的百分数是较昂贵的饭店（42.4%的饭店餐价是30～39美元，33.4%的饭店餐价是40～49美元）。因此，我们继续看到较昂贵的餐价与较高质量等级的饭店相联系。

交叉分组表被广泛地用于考察两个变量间的关系。在实践中，许多统计研究的最终报告包括大量的交叉分组

表。在洛杉矶饭店调查中，交叉分组表是基于一个分类变量（质量等级）和一个数量变量（餐价）编制的。当两个变量都是分类变量或数量变量时，也可以编制交叉分组表。然而，当使用数量变量时，我们首先必须对变量值划分组别。例如，在饭店这个例子中，我们将餐价划分为四个组（10～19 美元，20～29 美元，30～39 美元和 40～49 美元）。

表 2-11 每一个质量等级类的行百分比

质量等级	餐价（美元）				
	10～19	20～29	30～39	40～49	总计
好	50.0	47.6	2.4	0.0	100
很好	22.7	42.7	30.6	4.0	100
优秀	3.0	21.2	42.2	33.4	100

2.3.2 辛普森悖论

我们常常合并或综合两个或以上的交叉分组表中的数据，生成一个汇总的交叉分组表，以显示两个变量的相关性。在这种情形下，从两个或多个单独的交叉分组表得到的结论与一个综合的交叉分组表数据得到的结论可能截然相反。依据综合和未综合数据得到的相反的结论被称为**辛普森悖论**（Simpson's paradox）。为了说明辛普森悖论，我们考虑一个分析两位法官在两个不同的法庭上判决的例子。

在过去的三年中，法官罗恩·勒基特和丹尼斯·肯德尔在民事庭和市政庭主持审理案件，他们判决的部分案件被提出上诉。上诉法庭对大多数上诉案件维持原来的判决，但也有部分判决被推翻。以两个变量：判决（维持或推翻）和法庭类型（民事庭或市政庭）为依据，对每位法官构建交叉分组表。然后，假设通过综合法庭类型数据将两个交叉分组表合并。综合后的交叉分组表包含两个变量：判决（维持或推翻）和法官（勒基特或肯德尔）。这个交叉分组表给出了两位法官的上诉案件被判决维持或推翻的数量。在下面的交叉分组表中给出了这些结果，括号中的数是其旁边数值的列百分数。

判决	法官		总计
	勒基特	肯德尔	
维持	129（86%）	110（88%）	239
推翻	21（14%）	15（12%）	36
总计（%）	150（100%）	125（100%）	275

观察列百分数，我们看到法官勒基特有 86% 的判决维持原来的判决，法官肯德尔有 88% 的判决维持原来的判决。从这个综合交叉分组表，我们可以认为法官肯德尔做得比较好，因为他的判决上诉后维持原来判决的比例比较高。但是，问题随之出现。

下面未综合的交叉分组表是勒基特和肯德尔在两类法庭审理的案件，括号中的数是其旁边数值的列百分数。

判决	法官勒基特		总计
	民事庭	市政庭	
维持	29（91%）	100（85%）	129
推翻	3（9%）	18（15%）	21
总计（%）	32（100%）	118（100%）	150

判决	法官肯德尔		总计
	民事庭	市政庭	
维持	90（90%）	20（80%）	110
推翻	10（10%）	5（20%）	15
总计（%）	100（100%）	25（100%）	125

从勒基特的交叉分组表和列百分数，我们看到他在民事庭审理的案件中有 91% 上诉案件维持原来的判决，在

市政庭审理的案件中有85%维持原来的判决。从肯德尔的交叉分组表和列百分数，我们看到他在民事庭审理的案件中有90%上诉案件维持原来的判决，在市政庭审理的案件中有80%维持原来的判决。因此，当我们未综合数据时，我们看到法官勒基特有更好的记录，因为法官勒基特的判决在两个法庭中维持原来的判决的比例更高。这个结果与我们综合两个法庭数据得到的结论相矛盾。在最初的交叉分组表中，表明法官肯德尔有较好的记录。依据综合和未综合数据得到的相反的结论说明了辛普森悖论。

最初的交叉分组表是综合两个法庭分开的交叉分组表中的数据得到的。注意，对两位法官来说，上诉案件被推翻的比例在市政庭比民事庭要高。因为法官勒基特审理的案件大多数在市政庭，因此综合后的数据偏向于法官肯德尔。可是，当我们观察两个法庭单独的交叉分组表时，清楚地显示了法官勒基特有较好的记录。因此，对最初的综合交叉分组表，我们看到法庭类型是一个隐藏的变量，当评价两位法官的记录时，它是不可忽视的变量。

由于辛普森悖论的可能性，应该认识到根据未综合还是综合交叉分组表数据得出的结论或解释有可能截然相反。在得出结论之前，你应该审查交叉分组表是综合形式还是未综合形式，以便提出较好的见解和结论。特别地，当交叉分组表包括综合数据时，你应该审查是否存在可能影响结论的隐藏变量，使得分开的或未综合交叉分组表提供不同的、可能更好的见解和结论。

练　习

方法

28. 下列是两个数量变量 x 和 y 的观测结果。

观测次数	x	y	观测次数	x	y
1	28	72	11	13	98
2	17	99	12	84	21
3	52	58	13	59	32
4	79	34	14	17	81
5	37	60	15	70	34
6	71	22	16	47	64
7	37	77	17	35	68
8	27	85	18	62	67
9	64	45	19	30	39
10	53	47	20	43	28

a. 以 x 为行变量，y 为列变量，编制数据的交叉分组表。对 x，使用 10～29，3～49 等分组，对 y，使用 40～59，60～79 等分组。

b. 计算行百分数。

c. 计算列百分数。

d. 如果 x 和 y 之间存在相关关系的话，将会是什么类型的关系？

应用

30. 再次参考第29题的家庭收入和教育水平的交叉分组表。

a. 计算列百分数，并指出百分数频数分布。家长高中未毕业的比例是多少？

b. 家庭收入在100 000美元及以上，家长有学士以上学历的比例是多少？家长有学士以上学历，家庭收入在100 000美元及以上的比例是多少？二者有何区别？

c. 比较家庭收入在“25以下”、“100及以上”和“总计”的百分数频数分布。评价家庭收入和家长的教育水平之间的关系。

32. 表2-12是一个由45只共同基金的信息组成的数据集，它们是2008年的晨星基金500的一部分，数据集中包含下列5个变量：

基金类型：标记为DE（国内股本）、IE（国际股本）和FI（固定收益）

资产净值（美元）：每份的收盘价

5年的平均回报率（%）：基金过去5年的平均年回报率

费用比率：每财政年度扣除的基金费用占资产的百分比

晨星评级：每只基金风险调整星级，晨星星级从最低的1星级到最高的5星级

a. 以基金类型为行，以5年的平均回报率为列，编制数据的交叉分组表，5年的平均回报率（%）按0～9.99，10～29.99，30～39.99，40～49.99和50～59.99分组。

b. 编制基金类型数据的频数分布。

c. 编制5年的平均回报率（%）数据的频数分布。

d. 交叉分组表是如何有助于编制（b）和（c）的频数分布的？

e. 关于基金类型和过去5年的平均回报率，你能得出什么结论？

表2-12 45只共同基金样本的财务数据

基金名称	基金类型	净资产值（美元）	5年的平均回报率（%）	费用率（%）	晨星评级
Amer Cent Inc & Growth Inv	DE	28.88	12.39	0.67	2星
American Century Intl. Disc	IE	14.37	30.53	1.41	3星
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4星
American Century Ultra	DE	24.94	10.88	0.99	3星
Ariel	DE	46.39	11.32	1.03	2星
Artisan Intl Val	IE	25.52	24.95	1.23	3星
Artisan Small Cap	DE	16.92	15.67	1.18	3星
Baron Asset	DE	50.67	16.77	1.31	5星
Brandywine	DE	36.58	18.14	1.08	4星
Brown Cap Small	DE	35.73	15.85	1.20	4星
Buffalo Mid Cap	DE	15.29	17.25	1.02	3星
Delafield	DE	24.32	17.77	1.32	4星
DFA U. S. Micro Cap	DE	13.47	17.23	0.53	3星
Dodge & Cox Income	FI	12.51	4.31	0.44	4星
Fairholme	DE	31.86	18.23	1.00	5星
Fidelity Contrafund	DE	73.11	17.99	0.89	5星
Fidelity Municipal Income	FI	12.58	4.41	0.45	5星
Fidelity Overseas	IE	48.39	23.46	0.90	4星
Fidelity Sel Electronics	DE	45.60	13.50	0.89	3星
Fidelity Sh-Term Bond	FI	8.60	2.76	0.45	3星
Fidelity	DE	39.85	14.40	0.56	4星
FPA New Income	FI	10.95	4.63	0.62	3星
Gabelli Asset AAA	DE	49.81	16.70	1.36	4星
Greenspring	DE	23.59	12.46	1.07	3星
Janus	DE	32.26	12.81	0.90	3星
Janus Worldwide	IE	54.83	12.31	0.86	2星
Kalmar Gr Val Sm Cp	DE	15.30	15.31	1.32	3星
Managers Freemont Bond	FI	10.56	5.14	0.60	5星
Marsico 21st Century	DE	17.44	15.16	1.31	5星
Mathews Pacific Tiger	IE	27.86	32.70	1.16	3星
Meridan Value	DE	31.92	15.33	1.08	4星
Oakmark I	DE	40.37	9.51	1.05	2星
PIMCO Emerg Mkts Bd D	FI	10.68	13.57	1.25	3星
RS Value A	DE	26.27	23.68	1.36	4星
T. Rowe Price Latin Am.	IE	53.89	51.10	1.24	4星
T. Rowe Price Mid Val	DE	22.46	16.91	0.80	4星
Templeton Growth A	IE	24.07	15.91	1.01	3星
Thornburg Value A	DE	37.53	15.46	1.27	4星
USAA Income	FI	12.10	4.31	0.62	3星
Vanguard Equity-Inc	DE	24.42	13.41	0.29	4星
Vanguard Global Equity	IE	23.71	21.77	0.64	5星
Vanguard GNMA	FI	10.37	4.25	0.21	5星
Vanguard Sht-Tm TE	FI	15.68	2.37	0.16	3星
Vanguard Sm Cp Idx	DE	32.58	17.01	0.23	3星
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4星

34. 参见表2-12中的数据。

a. 以5年的平均回报率（%）为行，资产净值为列，编制数据的交叉分组表。5年的平均回报率以0为起点，5为组距分组；资产净值以0为起点，10为组距分组。

b. 如果变量间存在关系，对它们之间的关系进行评价。

2.4　用图形显示方式汇总两个变量的数据

在上一节，我们介绍了如何使用交叉分组表汇总两个变量的数据，并解释变量间的关系。在更多的情形中，图形显示对识别数据的模式和趋势更有效。

本节，我们介绍探索两个变量间关系的各种图形显示。用创造性的方法显示数据可以产生强大的洞察力，并使我们将依据“常识推理”变成直观的比较、对比，并识别模式。我们从散点图和趋势线的讨论开始。

2.4.1　散点图和趋势线

散点图（scatter diagram）是对两个数量变量间关系的图形表述，**趋势线**（trendline）是显示相关性近似程度的一条直线。作为一个实例，考虑旧金山一个音像设备商店的广告次数与销售额之间的关系。该商店在过去的三个月内有10次利用周末电视广告来促销。管理人员想证实广告播出次数和下一周商店销售额之间是否有关系。在表2-14中给出了10周销售额的样本数据（单位：100美元）。

图2-7是表2-14中数据的散点图和趋势线[⊖]。广告次数（x）显示在横轴上，销售额（y）显示在纵轴上。对第一周，$x=2$ 和 $y=50$，在散点图上按这两个坐标画出该点。用相同的方法画出其他9周的点。注意，有两周做了一次广告，有两周做了两次广告，依此类推。

表2-14　音像设备商店的样本数据

周	广告次数 x	销售额（100美元）y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

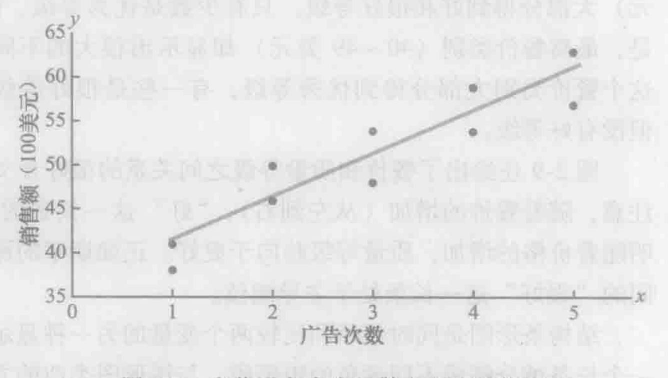


图2-7　音像设备商店的散点图和趋势线

绘制好的散点图（见图2-7）表明，广告次数和销售额之间存在正相关关系。较高的销售额与较高的广告次数相联系。因为所有的点并不在一条直线上，所以这种关系是不完全的。然而，这些点的分布模式和趋势线表明，整体关系是正相关的。

在图2-8中给出了一些一般的散点图模式和它们所显示的关系类型。左图描绘了正相关关系，与广告次数和销售额例子相似；而在中间的图中，散点图显示变量间没有明显的关系；在右图中，显示了负相关关系，即随着 x 增加，y 趋于减少。

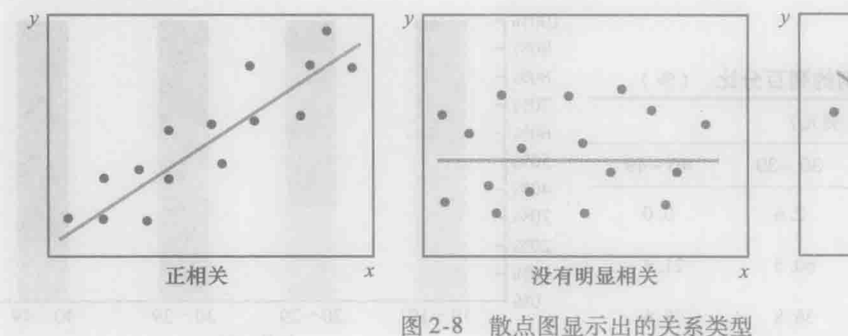
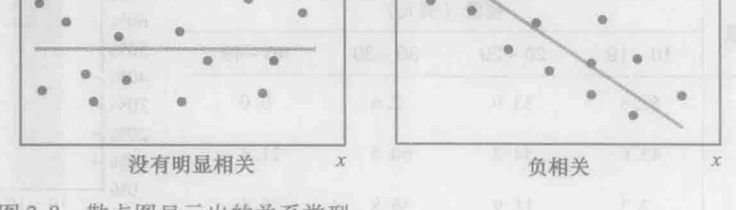



图2-8　散点图显示出的关系类型

⊖ 趋势线方程是 $y=36.15+4.95x$。趋势线的斜率是4.95，y 轴截距（趋势线与 y 轴的交点）是36.15。在第14章学习简单线性回归时，我们将详细讨论线性趋势线的斜率和 y 轴截距。

2.4.2 复合条形图和结构条形图

在第2.1节，我们提到条形图是描绘已汇总的分类型数据频数分布、相对频数分布或百分数频数分布的图形显示。复合条形图和结构条形图是基本条形图的扩展，其用于显示和比较两个变量。将两个变量显示在同一张图上，我们可以更好地了解变量间的关系。

复合条形图（side-by-side bar chart）是对已汇总的多个条形图同时显示的一种图形显示方式。为了说明复合条形图的绘制，回忆位于洛杉矶地区的300家饭店组成的一个样本，其质量等级和餐价数据的应用。质量等级是一个分类变量，等级类别有好、很好和优秀；餐价是一个数量变量，变化范围从10~49美元。如表2-10所示的交叉分组表表明餐价数据被分成四个组：10~19美元、20~29美元、30~39美元和40~49美元。我们将利用这些组来绘制复合条形图。

图2-9是饭店数据的复合条形图。每一个长条的灰度指明质量等级。在横轴上将每一个长条高度延伸，到达每一个餐价类别产生的质量等级的频数点上。每一个餐价类别的质量等级的频数彼此相邻，使我们迅速确定如何评价一个特殊的餐价类别。我们看到最低餐价类别（10~19美元）大部分得到好和很好等级，只有少数是优秀等级。但是，最高餐价类别（40~49美元）却显示出很大的不同。这个餐价类别大部分得到优秀等级，有一些是很好等级，但没有好等级。

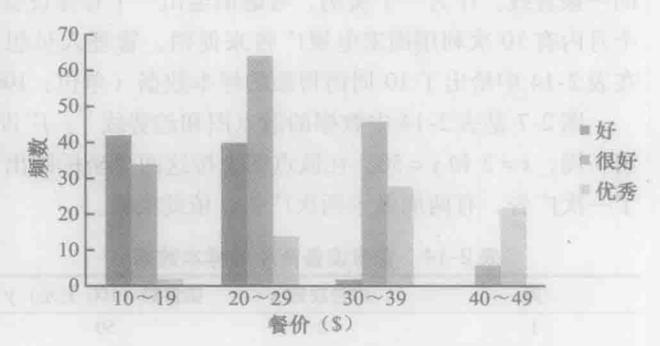


图2-9 质量等级和餐价数据的复合条形图

图2-9还给出了餐价和质量等级之间关系的很好含义。注意，随着餐价的增加（从左到右），“好”这一长条的高度在减少，而“优秀”这一长条的高度在增加。这表明随着价格的增加，质量等级趋向于更好。正如期望的那样，很好等级在中间价格类别更为突出，因为条形图中间的“很好”这一长条处于主导地位。

结构条形图是同时显示和比较两个变量的另一种显示方式。**结构条形图**（stacked chart）是一种条形图，其每一个长条被分解成不同颜色的矩形段，与饼形图类似的方式显示每一组的相对频数。为了说明结构条形图，我们将使用质量等级和餐价数据汇总的交叉分组表，如表2-10所示。

我们将表2-10中的某一列的每一个元素除以该列的总计数，可以将频数数据转化为列百分比。例如，在78家餐价在10~19美元的饭店中，有42家饭店具有好的质量等级。换句话说，78家饭店中有（42/78）×100或53.8%具有好的等级。表2-15给出了每一个餐价类别的列百分比。利用表2-15中的数据，我们绘制结构条形图，如图2-10所示。由于结构条形图以百分比为依据，因此图2-10甚至比图2-9更清楚地显示变量间的关系。当我们从最低价格组（10~19美元）到最高价格组（40~49美元）移动时，“好”长条的高度在减少，而“优秀”长条的长度在增加。

表2-15 每一个餐价类别的列百分比 （%）

质量等级	餐价（美元）			
	10~19	20~29	30~39	40~49
好	53.8	33.9	2.6	0.0
很好	43.6	54.2	60.5	21.4
优秀	2.3	11.9	36.8	78.6
总计	100.0	100.0	100.0	100.0

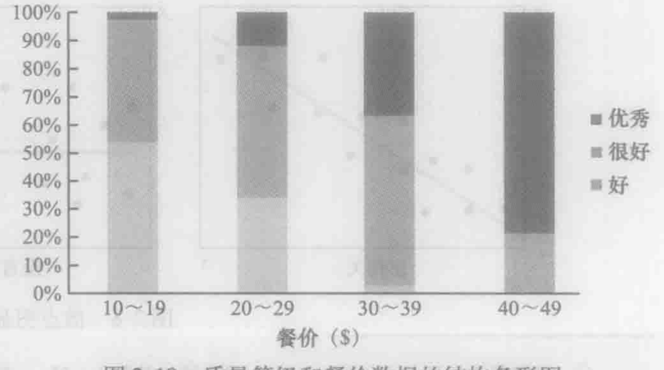


图2-10 质量等级和餐价数据的结构条形图

注释和评论

1. 时间序列是一个变量在连续时点或连续时期上测量的观测值的序列。以横轴表示时间值，纵轴表示时间序列值的散点图，在时间序列分析中称为时间序列图。我们将在第17章讨论时间序列图，以及如何分析时间序列数据。
2. 结构条形图也可以用频数而不是百分数频数来显示。在这种情形种，每个长条的不同颜色段得到长条的整体总数，而不是整体百分数。

练习

方法

☆36. 下列是两个数量变量 x 和 y 的20次观测结果。

观测次数	x	y	观测次数	x	y
1	−22	22	11	−37	48
2	−33	49	12	34	−29
3	2	8	13	9	−18
4	29	−16	14	−33	31
5	−13	10	15	20	−16
6	21	−28	16	−3	14
7	−13	27	17	−15	18
8	−23	35	18	12	17
9	14	−5	19	−20	−11
10	3	−3	20	−7	−22

a. 绘制出 x 和 y 之间的关系的散点图。

b. 如果 x 和 y 之间有明显的关系，将会是什么类型的关系？

38. 下面是由两个分类变量 x 和 y 的数据汇总的交叉分组表，变量 x 取值低、中或高，第二个变量 y 取值是或否。

x	y 是	y 否	总计
低	20	10	30
中	15	35	50
高	20	5	25
总计	55	50	105

a. 计算行百分数。

b. 以 x 为横轴，绘制百分数频数的结构条形图。

应用

40. 基于1981～2010年的数据，目前结果网站列出了美国51个主要城市的年平均最高气温、最低气温（单位：华氏度）和年平均降雪量（单位：英寸）。数据存在光盘中名为Snow的文件里。例如，俄亥俄州哥伦布市的平均最低气温是44度，年平均降雪量为27.5英寸。

a. 以年平均最低气温为横轴，年平均降雪量为纵轴，绘制散点图。

b. 这两个变量之间存在关系吗？

c. 以散点图为依据，对看似不寻常的数据点进行评价。

42. 智能手机是集互联网、照相、音乐和视频功能于一体的先进手机（The Pew Research Center，Internet & American Life Project，2011年）。下面的调查结果显示了智能手机拥有者的年龄。

年龄组	智能手机（%）	其他手机（%）	没有手机（%）
18～24	49	46	5
25～34	58	35	7
35～44	44	45	11
45～54	28	58	14
55～65	22	59	19
65＋	11	45	44

a. 以变量年龄组为横轴，绘制结构条形图，以显示上述调查数据的手机拥有者的类型。

b. 评论年龄和智能手机拥有者之间的关系。

c. 如果到了2021年，你期待这个调查的结果有怎样的不同？

2.5 数据可视化：创建有效图形显示的最佳实践

数据可视化是用于描述汇总和表述一个数据集信息图形显示的术语。数据可视化的目的是尽可能有效和清晰地传递数据的重要信息。本节，我们给出创建有效图形显示的指导准则，讨论如何根据研究的目的选择恰当的显示类型，说明数据仪表板的用途，以及展示辛辛那提动植物园如何利用数据仪表板技术改进决策。

2.5.1 创建有效的图形显示

表2-16中的数据给出了Gustin化学公司去年在美国销售地区的预测或计划销售额（单位：1 000美元）和实际销售额（单位：1 000美元）。注意，有两个数量变量（计划销售额和实际销售额）和一个分类变量（销售地区）。假设我们想建立一个能使Gustin化学公司管理层视觉化了解每个销售地区相对于计划销售额以及所有销售地区的销售业绩的图形显示。

图2-11是计划销售额与实际销售额数据的复合条形图。注意，这张条形图多么容易比较一个地区以及所有地区的计划与实际销售额。这个图形显示很简单，包含一个标题（也是标记）以及用不同颜色代表的两类销售额。注意，纵轴的尺度从0开始。四个销售地区分隔开，以表明它们是不同的，而为了便于每个地区内的比较，计划和实际销售额并列。在图2-11的复合条形图中，很容易看出西南地区的计划和实际销售额都是最低的，西北地区实际销售额略高于计划销售额。

表2-16　各销售地区的计划和实际销售额

销售地区	计划销售额（1 000美元）	实际销售额（1 000美元）
东北	540	447
西北	420	447
东南	575	556
西南	360	341

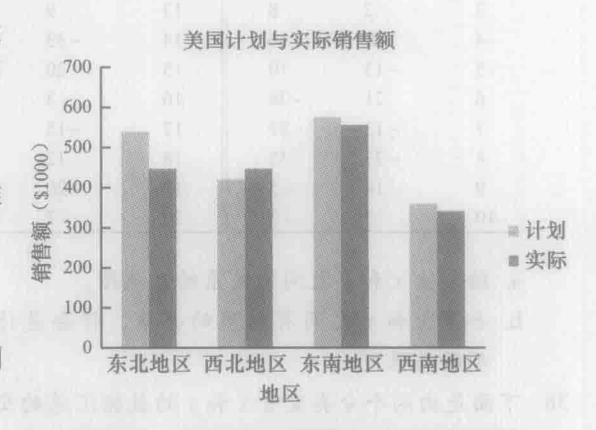


图2-11　计划和实际销售额的复合条形图

创建有效的图形显示是科学也是艺术。遵循如下的一般性准则，可以增强你的显示有效地表述数据中重要信息的可能性。

- 给予图形显示一个清晰、简明的标题。
- 使图形显示保持简洁，当能用二维表示时不要用三维表示。
- 每个坐标轴有清楚的标记，并给出测量的单位。
- 如果使用颜色来区分类别，要确保颜色是不同的。
- 如果使用多种颜色或线型，用图例来标明时，要将图例靠近所表示的数据。

2.5.2 选择图形显示的类型

本章我们讨论了各种图形显示，包括条形图、饼形图、打点图、直方图、茎叶显示、散点图、复合条形图和结构条形图。为了某种目的而绘制这些显示类型中的一种。为了给出选择合适的图形显示类型的指导准则，现在我们给出了根据目的分类的图形显示类型的汇总。我们注意到有些图形显示类型可有效地用于多个目的。

用于展示数据分布的图形显示

- 条形图　用于展示分类型数据的频数分布和相对频数分布
- 饼形图　用于展示分类型数据的相对频数分布和百分数频数分布
- 打点图　用于展示数值型数据在整个数据范围内的分布
- 直方图　用于展示数值型数据在一个区间组集合上的频数分布
- 茎叶显示　用于展示数值型数据的等级顺序和分布形态

用于进行比较的图形显示

- 复合条形图　用于两个变量的比较
- 结构条形图　用于比较两个分类变量的相对频数和百分数频数

用于展示相关关系的图形显示

- 散点图　用于展示两个数量变量的相关关系
- 趋势线　用于近似散点图中数据的相关关系

2.5.3 数据仪表板

数据可视化工具使用最广泛的一种是**数据仪表板**[㊀]（data dashboard）。如果你驾驶一辆汽车，你已经熟悉数据仪表板的概念。在一辆汽车中，汽车仪表板包括提供重要信息的测量仪表和其他直观显示，这些重要信息对驾驶一辆汽车非常重要。例如，用于显示车速、燃油油量、发动机温度和机油油量的仪表是确保安全和有效驾驶汽车的关键。在一些新款汽车中，为了给司机提供更有效的显示，这个信息甚至直观地显示在挡风玻璃上。对制定管理决策，数据仪表板扮演着相似的角色。

数据仪表板是一个直观显示的集合，这个直观显示用易于阅读、了解和解释的方式组织和表述用于监控公司或机构业绩的信息。就像车速、燃油油量、发动机温度和机油油量是监控汽车的重要信息一样，每一种行业有需要监控评估公司业绩状况的关键表现指标（KPIs）[㊁]。库存存货、日销售额、准时交货的比例和每季度销售收入都是 KPIs 的例子。数据仪表板应及时提供 KPIs 的（从各种可能来源）汇总信息，这对使用者很重要，而且应该做，以某种方式告诉它的使用者胜于控制它的使用者。

为了说明在决策中数据仪表板的使用，我们将讨论 Grogan 石油公司的一个应用。Grogan 石油公司在得克萨斯州的奥斯汀（其总部所在）、休斯敦和达拉斯这三个城市设有办事处。Grogan 石油公司位于奥斯汀办事处的信息技术呼叫中心处理员工关于软件、互联网和电子邮件问题等与计算机相关问题的呼叫。例如，如果一名在达拉斯的 Grogan 石油公司员工有计算机软件问题，则这名员工可以联络信息技术呼叫中心申请援助。

绘制监控呼叫中心表现的数据仪表板如图 2-12 所示。数据仪表板结合几种显示来监控呼叫中心的 KPIs，展示的是从上午 8 点开始的当前班次数据。左上角的结构条形图展示之前每一类问题（软件、互联网和电子邮件）的呼叫次数。条形图说明本班次最初的几个小时内呼叫次数比较多，关于电子邮件问题的呼叫次数随时间推移呈现递减趋势，关于软件问题的呼叫次数在 10 点最多。仪表板右上角的饼形图显示呼叫中心员工在每一类问题或没有呼叫工作（空闲）所用时间的比例。这些图对确定最佳人员配置水平都是重要显示。例如，以空闲时间的比例作为度量，了解呼叫成分比例和强调系统状况如何，可以帮助信息技术经理确信有足够的有专业技能的呼叫中心员工可供使用。

饼形图下面的复合条形图展示了 Grogan 石油公司每一个办事处的每一类问题的呼叫次数。这使得信息技术经理通过位置迅速识别是否存在某种特殊类型的问题。例如，图上显示出奥斯汀办事处在电子邮件问题上有相对高的数量。如果这些问题的原因能很快被识别出来，则大多数问题可能很快就被解决。同时，我们注意到，软件问题相对高的数量来自达拉斯办事处。这里较高的呼叫次数仅仅是由于达拉斯办事处刚刚安装了新软件，导致对信息技术中心的呼叫较多。由于上周达拉斯办事处已将此事提醒信息技术经理，信息技术经理知道来自达拉斯办事处的呼叫次数将会增加，可以增加人员配置水平以应对预期的呼叫次数的增加。

对每一个超过 15 分钟之前收到的未被解决的案例，数据仪表板中间左侧的条形图展示了这些未被解决的案例中每一个的时间长度。条形图使得 Grogan 石油公司迅速监控主要问题案例并决定是否需要额外资源来解决它们。最差的案例 T57 已超过 300 分钟未被解决，它实际上是从上一个班次遗留下来的。最后，在底部的直方图展示了当前班次所有被解决的案例所用时间的分布。

Grogan 石油公司的数据仪表板说明了正在运行水平中仪表板的使用。数据仪表板实时更新，用于如人员配置水平的运行决策。数据仪表板也可用于战术和战略水平管理。例如，一位物流经理为了其第三方运输的实时性能和成本而监控 KPIs，这有助于如运输方式和运输工具的选择等战术决策。在最高水平，更具战略性的数据仪表板将使上级管理部门通过监控更加综合的财务、服务水平和利用信息能力，迅速评估公司的财务状况。

前面讨论的好的数据可视化准则适用于数据仪表板中单个图，也适用于整个仪表板。除了这些准则之外，重要的是最大限度地减少屏幕滚过的需要、避免不必要的颜色使用和三维显示，以及图与图之间使用边框以提高可读性。作为单独的图，简单的永远是最好的。

㊀ 数据仪表板也称为数字仪表板。

㊁ 关键表现指标有时也称为关键表现测度（KPMs）。

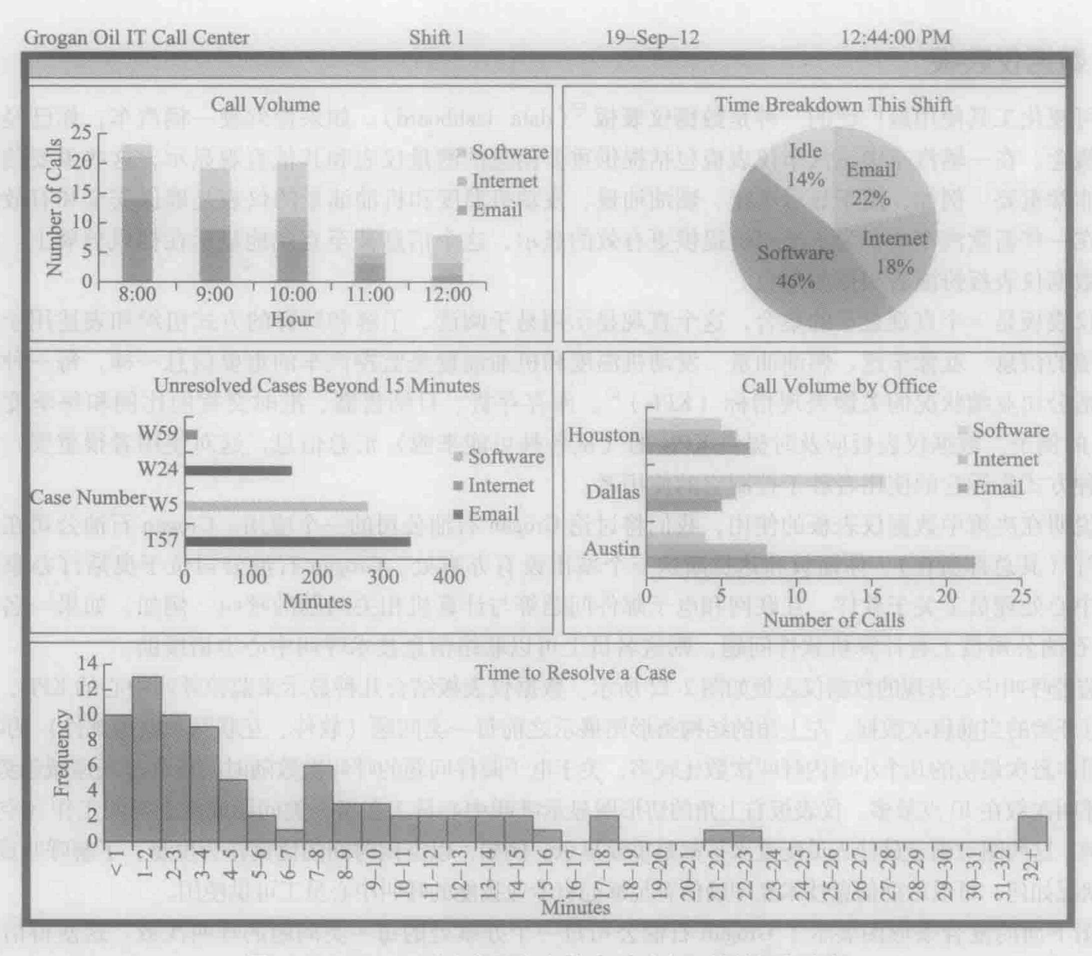


图 2-12 Grogan 石油公司信息技术呼叫中心的数据仪表板

2.5.4 实践中的数据可视化：辛辛那提动植物园㊀

位于辛辛那提市俄亥俄的辛辛那提动植物园是世界上第二古老的动物园。为了通过更多的数据驱动提高决策，管理者决定需要整合他们各个方面的业务，并规定非技术经理和主管用直观的方式更好地了解他们的数据。一个复杂的因素是当时动物园很忙，管理者期望与游客现场互动、检查运营、预计正在出现或即将发生的问题。因此，在实时基础上，能监控发生的情况是决定做什么的关键因素。动物园的管理者得出处理这类问题需要数据可视化战略的结论。

由于具有使用方便、实时更新能力和 iPad 兼容性，辛辛那提动植物园决定利用 IBM Cognos 的先进数据可视化软件来实行数据可视化战略。利用这个软件，辛辛那提动植物园开发了如图 2-13 所示的数据仪表板，它能使动物园的管理者追踪下面的关键表现指标：

- 项目分析（动物园内各处的销售量和销售额）
- Geo 分解（利用游客每日在动物园各处所用时间的分布图及显示）
- 顾客支出
- 收银员销售业绩
- 销售额和参观者资料与天气模式
- 动物园忠诚奖励计划的业绩

㊀ 作者感谢提供了本应用的辛辛那提动植物园的 John Lucas。

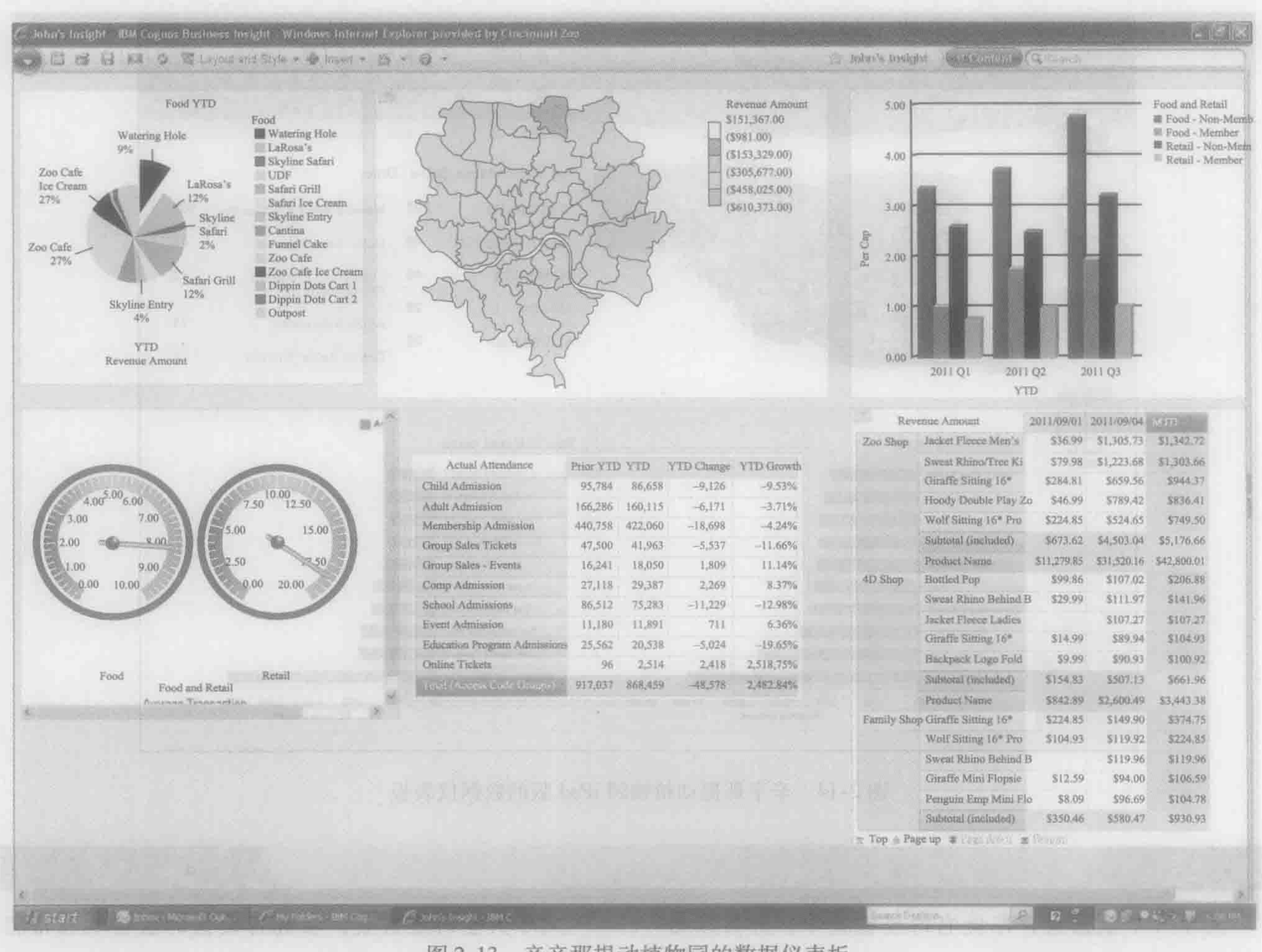


图 2-13　辛辛那提动植物园的数据仪表板

还开发的 iPad 移动应用能使动物园的管理者走到户外，在实时基础上仍然看到和预计到发生的情况。辛辛那提动物园的 iPad 版数据仪表板如图 2-14 所示，它可使管理者获取如下信息：

- 实时的参观者资料，包括游览动物园的游客“类型”
- 展示动物园内项目销售预测结果的实时分析
- 动物园游客所在位置的实时分布表示

获取如图 2-13 和图 2-14 所示的数据，可使动物园的管理者在动物园内人员配置水平方面、在天气和其他条件基础上需要储备的项目方面、在区域人口统计基础上如何更好地实现其广告目标方面作出更好的决策。

数据可视化的作用给动物园带来了显著的效果。在使用的第一年，系统直接负责的收入增长超过 500 000 美元，增加了动物园的游客人数，提高了顾客的服务水平和节约了成本。

注释和评论

1. 大多数软件适用于数据可视化。其中较为流行的软件包是 Cognos，JMP，Spotfire 和 Tableau。
2. 雷达图和气泡图是显示多个变量间关系的另外两种常用图形。但是，许多数据可视化专家建议不要使用这些图，原因是它们过于复杂。建议使用如条形图和散点图这些较为简单的显示图。
3. 地理数据可视化的非常强大工具是地理信息系统（GIS）。GIS 在地图上使用颜色、符号和标题来帮助我们了解变量的地理分布。例如，公司对设法确定新分布中心位置感兴趣，希望这个新分布中心能更好地了解公司产品需求量在全美的变动情况。可以使用 GIS 在地图上用红色区域表示高需求、蓝色区域表示低需求，没有颜色的区域表示未销售产品地区。越靠近红色高需求地区的位置越可能是下一步考虑好的候选地点。

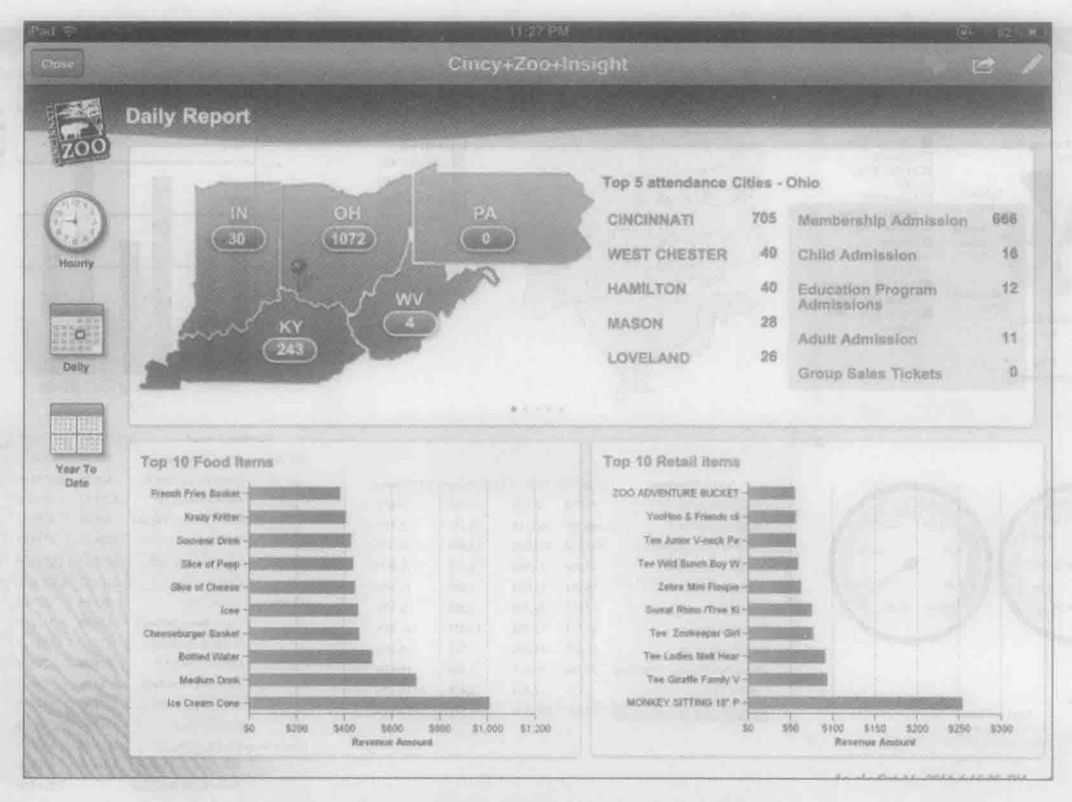


图2-14　辛辛那提动植物园 iPad 版的数据仪表板

小　结

对于一个数据集，即使它的规模适中，对其原始形式直接解释往往是困难的。表格法和图形法提供了组织和汇总数据的方法，使人们能够揭示出数据的特征模式，并能更容易地解释数据。频数分布、相对频数分布、百分数频数分布、条形图以及饼形图是用表格和图形汇总分类型数据的方法。频数分布、相对频数分布、百分数频数分布、直方图、累积频数分布、累积相对频数分布、累积百分数频数分布和茎叶显示是汇总数量型数据的方法。

交叉分组表是用表格汇总两个变量数据的方法。散点图是显示两个数量变量之间关系的图形方法。我们还展示了可以用于显示和比较两个分类变量的复合条形图和结构条形图，它们只是基本条形图的扩展。我们还讨论了创建有效图形显示的准则和如何选择最适合的显示类型。介绍了数据仪表板，说明如何构建一个用易于阅读、了解和解释的方式组织和表述用于监控公司或机构业绩的直观显示集合。图2-15是本章介绍的表格法和图示法的总结。

对于大型的数据集，在进行数据的表格和图形汇总时需要计算机软件包的帮助。在本章末尾的附录中，我们将说明如何使用 Minitab、Excel 和 StatTools 来达到这一目的。

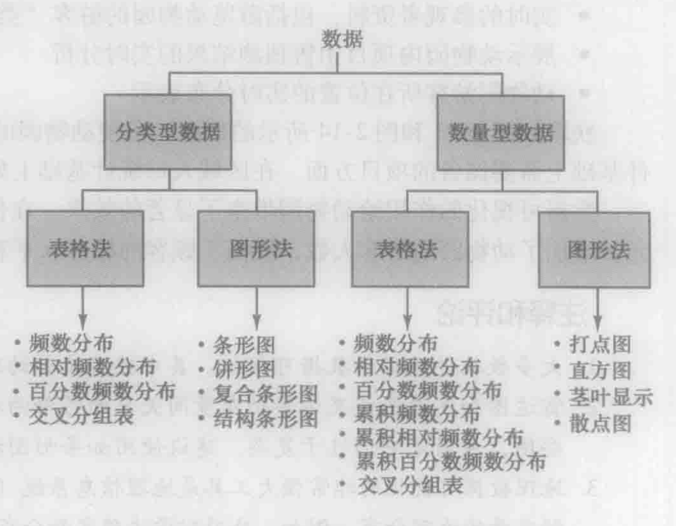


图2-15　汇总数据的表格法和图形法

关键术语

categorical data 分类型数据 用来识别相似项目类型的标记或名称。

quantitative data 数量型数据 表示大小或多少的数值。

data visualization 数据可视化 用于描述汇总和表述一个数据集信息的图形显示的效用的术语。

frequency distribution 频数分布 一种数据的表格汇总方法，表示在几个互不重叠组别中，每一组数据值的个数（频数）。

relative frequency distribution 相对频数分布 一种数据的表格汇总方法，表示在几个互无重叠组别中，每一组数据值个数的分数或所占比例。

percent frequency distribution 百分数频数分布 一种数据的表格汇总方法，表示在几个互不重叠组别中，每一组数据值个数所占的百分数。

bar graph 条形图 一种图形方法，用来描述已被汇总成频数分布、相对频数分布或百分数频数分布的分类型数据。

pie graph 饼形图 一种汇总数据的图形，该方法的依据是把一个圆细分为若干个扇形，使得每一组的相对频数与一个扇形相对应。

class midpoint 组中值 下组限和上组限的中间值。

dot plot 打点图 用横轴上每个数值上方的点的个数来汇总数据的一种图形方法。

histogram 直方图 一种描述数量型数据的频数分布、相对频数分布或百分数频数分布的图形方法，组宽放置在横轴上，频数、相对频数或百分数频数放置在纵轴上。

cumulative frequency distribution 累积频数分布 一种数量型数据的表格汇总方式，表示小于或等于每一组上组限的数据值的个数。

cumulative relative frequency distribution 累积相对频数分布 一种数量型数据的表格汇总方式，表示小于或等于每一组上组限的数据值的分数或比例。

cumulative percent frequency distribution 累积百分数频数分布 一种数量型数据的表格汇总方式，表示小于或等于每一组上组限的数据值的百分数。

stem-and-leaf display 茎叶显示 一种同时用于展示数据的等级排序和分布形态的图形显示。

crosstabulation 交叉分组表 一种对两个变量的数据进行的表格汇总，其中一个变量的组用行来描述，另一个变量的组用列来描述。

Simpson's paradox 辛普森悖论 从两个或两个以上单独的交叉分组表得到的结论可能与将数据综合成一个单一交叉分组表得出的结论截然相反。

scatter diagram 散点图 两个数量变量之间关系的图形表示，其中一个变量用横轴表示，另一个变量用纵轴表示。

trendline 趋势线 表示两个变量之间近似关系的一条直线。

side-by-side bar chart 复合条形图 描绘多个条形图同时显示的一种图形显示方式。

stacked chart 结构条形图 一种条形图，其每一个长条被分解成不同颜色的矩形段，与饼形图类似的方式显示每一组的相对频数。

data dashboard 数据仪表板 一个用易于阅读、了解和解释的方式组织和表述用于监控公司或机构业绩的直观显示集合。

重要公式

相对频数

$$\frac{\text{组频数}}{n} \tag{2-1}$$

近似组宽

$$\frac{\text{数据最大值}-\text{数据最小值}}{\text{组数}} \tag{2-2}$$

补充练习

44. 每年大约有150万名高中生参加学生能力倾向测试（SAT），近80%没有公开招生政策的学院和大学，在作出录取决定时用到了SAT分数（College Board，2009年3月）。目前SAT形式有三部分：阅读理解、数学和写作，SAT分数是三部分分数之和，满分是2 400分。一个SAT总分样本如下：

1 665	1 525	1 355	1 645	1 780
1 275	2 135	1 280	1 060	1 585
1 650	1 560	1 150	1 485	1 990
1 590	1 880	1 420	1 755	1 375
1 475	1 680	1 440	1 260	1 730
1 490	1 560	940	1 390	1 175

a. 编制频数分布和绘制直方图，第一组下限是800，组宽是200。

b. 对分布的形态进行评价。

c. 基于图形和表格汇总，关于SAT分数，还能得出什么观测结果？

46. 美国各州的人口数据如下（单位：100万人）（《世界年鉴》，2012年）。

州	人口数	州	人口数	州	人口数
亚拉巴马州	4.8	路易斯安那州	4.5	俄亥俄州	11.5
阿拉斯加州	0.7	缅因州	1.3	俄克拉何马州	3.8
亚利桑那州	6.4	马里兰州	5.8	俄勒冈州	4.3
阿肯色州	2.9	马萨诸塞州	6.5	宾夕法尼亚州	12.7
加利福尼亚州	37.3	密歇根州	9.9	罗得岛州	1.0
科罗拉多州	5.0	明尼苏达州	5.3	南卡来罗纳州	4.6
康涅狄格州	3.6	密西西比州	3.0	南达科他州	0.8
特拉华州	0.9	密苏里州	6.0	田纳西州	6.3
佛罗里达州	18.8	蒙大拿州	0.9	得克萨斯州	25.1
佐治亚州	9.7	内布拉斯加州	1.8	犹他州	2.8
夏威夷州	1.4	内华达州	2.7	佛蒙特州	0.6
爱达荷州	1.6	新罕布什尔州	1.3	弗吉尼亚州	8.0
伊利诺伊州	12.8	新泽西州	8.8	华盛顿州	6.7
印第安纳州	6.5	新墨西哥州	2.0	西弗吉尼亚州	1.9
艾奥瓦州	3.0	纽约州	19.4	威斯康星州	5.7
堪萨斯州	2.9	北卡罗来纳	9.5	怀俄明州	0.6
肯塔基州	4.3	北达科他州	0.7		

a. 以2.5（单位：100万人）为组宽，编制频数分布、百分数频数分布和绘制直方图。

b. 在分布中呈现出偏态吗？请解释。

c. 关于50个州的人口，你能得出什么观测结果？

48. 商业改进局经常收到消费者的投诉举报。在2011年，商业改进局收到投诉最多的行业是银行、有线和卫星电视公司、收债机构、移动电话运营商和新车经销店。一个200次投诉的样本结果存在名为BBB的文件中。

a. 展示分行业的投诉频数和百分数频数。

b. 绘制百分数分布的条形图。

c. 哪个行业的投诉次数最多？

d. 对投诉的百分数频数分布进行评论。

50.《金融时报》/哈里斯民意调查中的一个问题是："你赞成还是反对对碳排放较高的汽车征收较高的税？"可能的回答有非常赞成、赞成大于反对、反对大于赞成和非常反对。在4个欧洲国家和美国调查了5 372个成年人，下面是他们回答的交叉分组表（Harris Interactive website，2008年2月27日）。

支持水平	国家					合计
	英国	意大利	西班牙	德国	美国	
非常赞成	337	334	510	214	214	1 617
赞成大于反对	370	408	355	327	327	1 871
反对大于赞成	250	188	155	275	275	1 135
非常反对	130	115	89	204	204	749
合计	1 087	1 045	1 109	1 020	1 020	5 372

a. 编制支持水平变量的百分数频数分布。你是否认为：民意调查的结果是赞成对碳排放较高的汽车征收较高的税？

b. 编制国家变量的百分数频数分布。

c. 欧洲国家成年人的支持水平是否显著不同于美国成年人的支持水平？请解释。

52. 辛辛那提煤气与电力公司对商业建筑进行了一项调查，询问使用的主要取暖燃料是什么，以及商业建筑的建设年份。得到的部分结果的交叉分组表如下。

建设年份	燃料类型				
	电力	天然气	石油	丙烷	其他
1973年及以前	40	183	12	5	7
1974～1979	24	26	2	2	0
1980～1986	37	38	1	0	6
1987～1991	48	70	2	0	1

a. 通过计算行总计和列总计完成交叉分组表。

b. 编制建设年份和燃料类型的频数分布。

c. 编制显示列百分数的交叉分组表。

d. 编制显示行百分数的交叉分组表。

e. 对建设年份和燃料类型之间的关系进行评论。

54. 参见表2-18的数据集。

a. 以成立年代为行变量，毕业率为列变量编制交叉分组表。成立年代以1600为起点，2000为终点，50为组宽进行分组；毕业率以35%为起点，100%为终点，5%为组宽进行分组。

b. 计算（a）中交叉分组表的行百分数。

c. 对这两个变量之间的关系进行评价。

56. 参见表2-18的数据集。

a. 绘制学费和毕业率之间关系的散点图。

b. 对这两个变量之间的关系进行评价。

58. 动物园将游客分为三类：会员、学校和普通。会员类别定义为缴年费支持动物园的游客。会员会收到动物园给予的某些优惠（如商品折扣或旅行计划）。学校类别包括日托、小学和中等学校的全体从业人员和学生；这些游客会得到一个折扣比例。普通类别包括其他所有游客。动物园注意到近期参观者有所下降。为了有助于更好地了解参观者与会员的关系，一名动物园的员工搜集了如下数据。

游客类别	参观者			
	2008	2009	2010	2011
普通	153 713	158 704	163 433	169 106
会员	115 523	104 795	98 437	81 217
学校	82 885	79 876	81 970	81 290
合计	352 121	343 375	343 840	331 613

a. 绘制全体参观者关于随时间的条形图。对数据中的趋势进行评论。

b. 以年份为横轴上的变量，绘制展示参观者游客类别的复合条形图。

c. 根据（a）和（b）的条形图，对动物园参观者发生的情况进行评论。

案例2-1 Pelican商店

Pelican商店是National Clothing的一个分支，是一家在全美范围内经营妇女服饰的连锁商店。最近，它举办了一项促销活动，向其他National Clothing商店的顾客赠送优惠券。在促销活动期内的某一天，Pelican商店抽取了100名持信用卡交易的顾客组成一个样本，搜集到的数据存在名为PelicanStores的文件中。表2-19是数据集的一部分。Proprietary Card付款方法是指使用National Clothing签账卡收费。使用优惠券购物的顾客定义为促销顾客，没有使用优惠券购物的顾客定义为普通顾客。因为优惠券不会派发给Pelican商店的普通顾客，管理者认为，持促销优惠券的顾客生产的销售额与其他顾客不同。当然，Pelican的管理者也希望促销顾客会继续在他的商店购物。

表2-19中的大多数变量不需要解释，但有两个变量需要稍做说明。

表2-19 Pelican商店100名持信用卡购物的顾客的样本数据

顾客	顾客类型	项目	净销售额	支付方法	性别	婚姻状况	年龄
1	普通	1	39.50	Discover	男	已婚	32
2	促销	1	102.40	Proprietary Card	女	已婚	36
3	普通	1	22.50	Proprietary Card	女	已婚	32
4	促销	5	100.40	Proprietary Card	女	已婚	28
5	普通	2	54.00	Master Card	女	已婚	34
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
96	普通	1	39.50	Master Card	女	已婚	44
97	促销	9	253.00	Proprietary Card	女	已婚	30
98	促销	10	287.59	Proprietary Card	女	已婚	52
99	促销	2	47.60	Proprietary Card	女	已婚	30
100	促销	1	28.44	Proprietary Card	女	已婚	44

（1）项目：购买商品的总件数

（2）净销售额：信用卡支付的总金额（单位：美元）

Pelican的管理者希望使用这些样本数据，以了解其顾客的基本情况并对使用优惠券的促销活动进行评估。

管理报告

使用描述统计的表格和图形方法来帮助管理部门建立顾客档案，并对促销活动进行评估。你的报告至少应该

包括以下内容：

1. 主要变量的百分数频数分布。
2. 条形图或饼形图，以显示顾客使用各种付款方式的购物数量。
3. 顾客类型（普通或促销）与净销售额的交叉分组表，对相似性与差异性进行评价。
4. 探索净销售额与顾客年龄关系的散点图。

案例2-2 电影业

电影业是一个竞争激烈的行业，每年有超过50个制片厂制作出300～400部新电影，每部电影商业上的成功差异很大。周末首映票房收入（单位：100万美元）、票房总收入（单位：100万美元）、放映电影的剧院数，以及电影放映的周数是衡量一部电影是否成功最常用的变量。2011年制作的100部大电影的样本数据存在光盘中名为2011Movies文件中（Box Office Mojo，2012年3月17日）。表2-20是文件中前10部电影的有关数据。

表2-20 前10部电影的业绩数据

电影	首映票房（100万美元）	票房总收入（100万美元）	剧院数	放映周数
哈利波特与死亡圣器第二部	169.19	381.01	4 375	19
变形金刚：月黑之时	97.85	352.39	4 088	15
暮光之城：破晓第一部	138.12	281.29	4 066	14
宿醉2	85.95	254.46	3 675	16
加勒比海盗：惊涛怪浪	90.15	241.07	4 164	19
速度与激情5	86.20	209.84	3 793	15
碟中谍4：幽灵协议	12.79	208.55	3 555	13
赛车总动员2	66.14	191.45	4 115	25
福尔摩斯：诡影游戏	39.64	186.59	3 703	13
雷神	65.72	181.03	3 963	16

管理报告

使用描述统计学的表格和图形方法来了解这些变量对一部电影成功有怎样的贡献。你的报告应该包括以下内容：

1. 四个变量中每个变量的表格与图形汇总，接着根据每个汇总告诉我们有关电影业的情况。
2. 探究票房总收入与周末首映票房收入之间的散点图，并讨论。
3. 探究票房总收入与剧院数之间的散点图，并讨论。
4. 探究票房总收入与在放映周数之间的散点图，并讨论。

第3章

描述统计学Ⅱ：数值方法

CHAPTER 3

实践中的统计

Small Fry 设计公司[㊀]

加利福尼亚州，圣安娜

Small Fry 设计公司成立于1997年，是一家设计和进口婴儿产品的公司，主要经营玩具和附属用品。公司的产品包括泰迪熊、悬挂饰物、音乐玩具、拨浪鼓和安全毯，以及设计注重颜色、材质和声音的高品质的柔软玩具。公司的产品在美国设计，在中国生产。

Small Fry 设计公司使用独立的销售代理，把产品分销给婴儿用品零售商、儿童用品及服饰商店、礼品店、大型百货商店和主要的邮购公司。目前，Small Fry 设计公司的产品遍布美国的1 000 多个零售批发商店。

在这家公司的日常运营中，现金流管理是最重要的经营活动之一。是否能够保证公司拥有足够的现金收入，以满足目前和未来的偿债义务，决定着公司业务的成败。现金流管理的一个关键因素是对应收账款的分析和控制。通过度量未付款发票的平均期限和资金数额，管理人员能够预测现金供应和监视应收账款状态的变化。公司设置了如下目标：未付款发票的平均期限不应超过45天；超过60天的未付款发票的资金数额不应超过所以应收账款总额的5%。

在最近对应收账款的汇总中，该公司使用了下面的描述统计量度来衡量未付款发票的期限：

平均数	40 天
中位数	35 天
众数	31 天

对这些统计量的解释表明，一张发票的平均数或平均期限是40天；中位数表明有一半的发票已经超过35天没有付款；最常见的发票期限是众数为31天，表明一张未付款发票最常见的时间长度是31天。统计汇总还显示出应收账款总额中只有3%超过60天。基于这些统计信息，管理人员可以感到满意，因为应收账款和收入现金流都处于控制之中。

在本章中，你将学到如何计算和理解 Small Fry 设计公司所使用过的一些描述统计量。除了平均数、中位数和众数外，你还将学到其他的描述统计量，比如极差、方差、标准差、百分位数和相关系数。这些数值测度将有助于读者对数据的理解和解释。

在第2章中，我们讨论了汇总数据的表格和图形方法。在本章中，我们将给出几种描述统计学的数值方法，它们提供了汇总数据的其他可选方法。

首先，我们对只包含一个变量的数据集建立数值汇总的度量方法。当数据集包含的变量不止一个时，可以对每个变量分别计算其相同的数值测度。然而，在有两个变量的情况下，我们还将建立变量间相互关系的度量。

本章我们将介绍位置、离散程度、形态和相关程度的数值度量。如果数据来自样本，计算的度量称为**样本统计量**（sample statistics）。如果数据来自总体，计算的度量称为**总体参数**（population parameters）。在统计推断中，样本统计量被称为是相应总体参数的**点估计量**（point estimator）。在第7章，我们将详细讨论点估计方法。

在本章的三个附录中，我们将演示如何使用 Minitab Excel 和 StatTools 来计算本章介绍的数值方法。

3.1 位置的度量

3.1.1 平均数

在位置的数值度量中，最重要的大概要数变量的**平均数**[㊁]（mean）或平均值了。平均数提供了数据中心位置的度量。如果数据来自某个样本，则平均数用 $\bar{x}$ 表示；如果数据来自某个总体，则平均数用希腊字母 μ 表示。

在统计公式中，我们习惯用 x_1 来表示变量 x 的第一个观测值，用 x_2 来表示变量 x 的第二个观测值，依此类推。一般地，用 x_i 来表示变量 x 的第 i 个观测值。对一个有 n 个观测值的样本，其样本平均数的计算公式如下：

㊀ 作者感谢为“实践中的统计”提供了本案例的 Small Fry 设计公司总裁约翰 A. 麦卡锡。

㊁ 平均数有时也称为算术平均数。

样本平均数[⊖]

$$\bar{x} = \frac{\sum x_i}{n} \tag{3-1}$$

在式（3-1）中，分子是 n 个观测值的数值之和，即

$$\sum x_i = x_1 + x_2 + \cdots + x_n$$

希腊字母 $\sum$ 是求和符号。

为了说明样本平均数的计算，由大学的5个班级组成一个样本，每个班级的学生人数是：

46　　54　　42　　46　　32

我们用符号 x_1，x_2，x_3，x_4 和 x_5 分别表示这5个班的学生人数。

$$x_1 = 46 \quad x_2 = 54 \quad x_3 = 42 \quad x_4 = 46 \quad x_5 = 32$$

因此，为计算样本平均数，我们有

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

班级人数的样本平均数是44人。

为了给出平均数直观表述，并展示其如何受到极端值的影响，考虑如图3-1所示的班级人数数据的打点图。将用于绘制打点图的横轴处理成一个又长又窄的平面，在这个平面上每个点都有相同的固定权重。平均数是为了平衡打点图我们放置在平面下方的支点或中心点。这与运动场上的跷跷板的工作原理相同，唯一不同的是跷跷板的中心点在中间，使得当一端上升时，另一端下降。在打点图上，我们根据点的位置来设置中心点。现在如果我们将最大值从54增大到114，考虑平衡会发生怎样变化。为了重新建立平衡，在新的打点图下，我们需要将支点向正的方向移动。为了确定我们支点移动的数值，我们只要计算修改后班级人数数据的样本平均数。

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 114 + 42 + 46 + 32}{5} = 56$$

因此，修改后班级人数数据的平均数为56，增加了12人。换句话说，为了建立平衡，在新的打点图下，我们需要将平衡点向右移动12个单位。

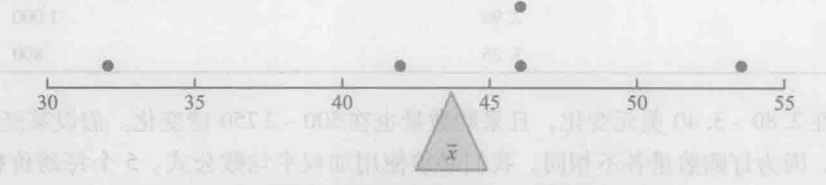


图3-1　平均数作为班级人数数据打点图的平衡中心

下面是计算样本平均数的另一个例子。假设某大学生就业指导办公室对一个商学院毕业生的样本进行问卷调查，以获取大学毕业生起始月薪的有关信息。表3-1给出了搜集到的数据，12名商学院毕业生的平均起始月薪的计算如下：

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12} = \frac{3\,850 + 3\,950 + \cdots + 3\,880}{12} = \frac{47\,280}{12} = 3\,940$$

表3-1　12名商学院毕业生样本的起始月薪数据

毕业生	起始月薪（美元）	毕业生	起始月薪（美元）
1	3 850	7	3 890
2	3 950	8	4 130
3	4 050	9	3 940
4	3 880	10	4 225
5	3 755	11	3 920
6	3 710	12	3 880

式（3-1）说明了有 n 个观测值的样本如何计算平均数。而计算总体平均数的公式也基本相同，但是我们使用不同的符号表示我们是对整个总体计算的。我们用 N 表示总体观测值的个数，用 μ 表

⊖ 样本平均数 $\bar{x}$ 是一个样本统计量。

示总体平均数。

总体平均数[1]

$$\mu = \frac{\sum x_i}{N} \tag{3-2}$$

3.1.2 加权平均数

在样本平均数和总体平均数的公式中，每个 x_i 都有相同的重要性或权重。例如，样本平均数的公式可以写成如下形式：

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1}{n}\sum x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}(x_1) + \frac{1}{n}(x_2) + \cdots + \frac{1}{n}(x_n)$$

这表明样本中每一个观测值都有权重 $1/n$。虽然在实践中这种情况最常见，但有时，计算平均数时会对每个观测值赋予显示其重要性的权重。以这种方式计算的平均数称作**加权平均数**（weighted mean）。加权平均数的计算公式如下：

加权平均数

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \tag{3-3}$$

式中，w_i——第 i 个观测值的权重。

当数据来自样本时，式（3-3）计算的是加权样本平均数。当数据来自总体时，式（3-3）计算的是加权总体平均数，并用 μ 取代 $\bar{x}$。

作为一个需要加权平均数的例子，考虑下面一个在过去 3 个月中 5 次购买原材料的样本。

购买批次	价格（美元/磅）	数量（磅）
1	3.00	1 200
2	3.40	500
3	2.80	2 750
4	2.90	1 000
5	3.25	800

注意，每磅价格在 2.80～3.40 美元变化，且采购数量也在 500～2 750 磅变化。假设某经理想知道原材料每磅平均价格的有关信息。因为订购数量各不相同，我们必须使用加权平均数公式。5 个每磅价格的数据分别为 $x_1 = 3.00$，$x_2 = 3.40$，$x_3 = 2.80$，$x_4 = 2.90$ 和 $x_5 = 3.25$。计算每磅价格的平均数时，要根据每个价格所对应的数量来加权。在本例中，五个权重分别为：$w_1 = 1\,200$、$w_2 = 500$、$w_3 = 2\,500$、$w_4 = 1\,000$ 和 $w_5 = 800$。利用式（3-3），计算加权平均数如下：

$$\bar{x} = \frac{1\,200 \times 3.00 + 500 \times 3.40 + 2\,750 \times 2.80 + 1\,000 \times 2.90 + 800 \times 3.25}{1\,200 + 500 + 2\,750 + 1\,000 + 800} = \frac{18\,500}{6\,250} = 2.96$$

于是，使用加权平均计算原材料每磅的平均价格为 2.96 美元。注意，如果不是使用加权平均数公式而是使用式（3-1）计算的话，就会得到带有误导性的结果。在这种情况下，5 个每磅价格的平均数为 $(3.00 + 3.40 + 2.80 + 2.90 + 3.25)/5 = 15.35/5 = 3.07$ 美元，夸大了每磅实际的平均价格。

在某个特殊的加权平均数的计算中，对权重的选择依赖于具体的应用。一个被大学生们所熟知的例子就是平均等级分（grade point average，GPA）的计算。在计算中，普遍使用的数值是：A 级是 4，B 级是 3，C 级是 2，D

[1] 样本平均数 $\bar{x}$ 是总体平均数 μ 的点估计。

级是1，F级是0。权重就是取得每一等级的学分数。本节末的第16题给出了计算GPA加权平均数的一个例子。在其他加权平均数的计算中，经常会使用诸如磅、美元、体积等数值作为权重。在任何情况下，当观测值的重要性变化时，分析人员必须选择能够最好地反映每个观测值重要性的权重，来计算加权平均数。

3.1.3 中位数

中位数（median）是对变量中心位置的另一种度量。将所有数据按升序（从小到大的顺序）排列后，位于中间的数值即为中位数。当观测值是奇数时，中位数就是位于中间的那个数值；当观测值是偶数时，则没有单一的中间数值。在这种情况下，我们遵循惯例，定义中位数为中间两个观测值的平均值。为了方便起见，我们重新定义中位数如下。

> **中位数**
>
> 将数据按升序（从小到大的顺序）排列。
>
> (a) 对奇数个观测值，中位数是中间的数值。
>
> (b) 对偶数个观测值，中位数是中间两个数值的平均值。

我们用上述定义来计算5个大学班级样本的班级人数的中位数。将这5个数值按升序排列如下：

32　42　46　46　54

由于 $n=5$ 是奇数，则中位数是中间的数值，因此，班级人数的中位数是46人。当我们按升序排列数据时，尽管数据中有两个观测值是46，但我们对每个观测值单独处理。

我们再来计算表3-1中商学院毕业生起薪的中位数。我们首先将数据按升序排列如下：

3 710　3 755　3 850　3 880　3 880　3 890　3 920　3 940　3 950　4 050　4 130　4 325

（3 890　3 920：中间两个值）

由于 $n=12$ 是偶数，我们找出中间的两个值：3 890和3 920。中位数就是它们的平均值。

$$\text{中位数}=\frac{3\,890+3\,920}{2}=3\,905$$

我们用于计算中位数的方法依赖于是有奇数个观测值还是有偶数个观测值。现在，让我们用12名商学院毕业生起始月薪来描述更具概念性和直观的方法。之前，我们首先将数据按升序排列如下：

3 710　3 755　3 850　3 880　3 880　3 890　3 920　3 940　3 950　4 050　4 130　4 325

一旦数据按升序排列，我们去掉一个极端大值和极端小值构成的对，直到没有更多的数值对可以去掉为止，但不能消除所有的数据。例如，在去掉最小观测值（3 710）和最大观测值（4 325）后，我们得到有10个观测值的新数据集。

~~3 710~~　3 755　3 850　3 880　3 880　3 890　3 920　3 940　3 950　4 050　4 130　~~4 325~~

然后，我们去掉其次剩余的最小值（3 755）和其次剩余的最大值（4 130），产生有8个观测值的新数据集。

~~3 710~~　~~3 755~~　3 850　3 880　3 880　3 890　3 920　3 940　3 950　4 050　~~4 130~~　~~4 325~~

继续这个过程，我们得到如下结果。

~~3 710~~　~~3 755~~　~~3 850~~　3 880　3 880　3 890　3 920　3 940　3 950　~~4 050~~　~~4 130~~　~~4 325~~

~~3 710~~　~~3 755~~　~~3 850~~　~~3 880~~　3 880　3 890　3 920　3 940　~~3 950~~　~~4 050~~　~~4 130~~　~~4 325~~

~~3 710~~　~~3 755~~　~~3 850~~　~~3 880~~　~~3 880~~　3 890　3 920　~~3 940~~　~~3 950~~　~~4 050~~　~~4 130~~　~~4 325~~

在这一点，没有更多的可以去掉且不能消除所有的数据是可能的。因此，中位数恰好是剩余两个值的平均值。当有偶数个观测值时，去掉过程将永远导致两个剩余值，这些值的平均值就是中位数。当有奇数个观测值时，去掉过程将永远导致一个最终值，这个值就是中位数。因此，无论观测值个数是奇数还是偶数，这种方法都适用。

虽然在度量数据的中心位置时，更常用的是平均数。但在某些情况下，使用中位数则更为合适。[⊖]平均数往往会受到异常大或异常小的数值影响。例如，假设某个毕业生（见表3-1）的起薪为每月10 000美元（也许他的家庭拥有这家公司）。如果我们把表3-1的最高起始月薪从4 325美元改为10 000美元，再重新计算平均数，则样本平均数将从3 940美元变为4 413美元。而中位数却没有改变，仍为3 905。因为3 890和3 920还是中间的两个值。当存在异常高的起薪情况时，中位数提供了比平均数更好的中心位置度量。我们通常说，在数据集含有异常值的情况下，中位数往往更适合于度量数据的中心位置。

3.1.4　几何平均数

几何平均数（geometric mean）是一种位置度量，它是 n 个数值乘积的 n 次方根。几何平均数记为 $\bar{x}_g$，其一般公式如下。

几何平均数

$$\bar{x}_g = \sqrt[n]{x_1 x_2 \cdots x_n} = (x_1 x_2 \cdots x_n)^{1/n} \tag{3-4}$$

几何平均数常常用于分析财务数据的增长率。在这种情形中，算术平均数或平均值会给出误导性结果。

为了说明几何平均数的用途，考虑一只共同基金过去10年间的年回报率或增长率，如表3-2所示。假设第1年年初在该基金投资100美元，我们想计算第10年年末的价值。我们首先计算该基金第1年年末的余额。由于第1年的年回报率为 -22.1%，则该基金第1年年末的余额将为

表3-2　共同基金数据的年回报率和增长因子

年	回报率（%）	增长因子
1	-22.1	0.779
2	28.7	1.287
3	10.9	1.109
4	4.9	1.049
5	15.8	1.158
6	5.5	1.055
7	-37.0	0.630
8	26.5	1.265
9	15.1	1.151
10	2.1	1.021

$$\begin{aligned}100 - 0.221 \times 100 &= 100 \times (1 - 0.221)\\ &= 100 \times 0.779 = 77.9(\text{美元})\end{aligned}$$

注意，在表3-2中0.779被称为增长因子[⊜]。这个结果表明，用第1年年初在该基金的投资额乘以第1年的增长因子，我们可以计算第1年年末的余额。

该基金第1年年末的余额77.9美元，现在变为第2年年初的余额。由于第2年的回报率为28.7%，则第2年年末的余额将为

$$77.9 + 0.287 \times 77.9 = 77.9 \times (1 + 0.287) = 77.9 \times 1.287 = 100.257\,3(\text{美元})$$

注意，1.287是第2年的增长因子。带入100×0.779或77.9（美元），我们看到第2年年末的余额为

$$100 \times 0.779 \times 1.287 = 100.257\,3(\text{美元})$$

换句话说，第2年年末的余额恰好是第1年年初的最初的投资额乘以两个增长因子的乘积。这个结果可以推而广之，得到第10年年末的余额为最初的投资额乘以所有增长因子的乘积。

$$\begin{aligned}&100 \times 0.779 \times 1.287 \times 1.109 \times 1.049 \times 1.158 \times 1.055 \times 0.630 \times 1.265 \times 1.151 \times 1.021\\ &= 100 \times 1.334\,493 = 133.449\,3(\text{美元})\end{aligned}$$

因此，第1年年初在该基金的100美元投资额，第10年年末价值将为133.449 3美元。注意，10个增长因子的乘积为1.334 493。因此，对第1年年初的任何数量的投资额，我们可以计算第10年年末的余额为最初的投资额乘以1.334 493。例如，第1年年初投资额为2 500美元，第10年年末价值将为2 500×1.334 493或大约3 336美元。

但是，过去10年间这项投资的平均年回报率或平均增长率是多少？让我们来看如何使用10个增长因子的几

⊖ 中位数作为位置的度量，经常用在年收入及资产价值数据的报告中，因为少数极端大的收入和资产价值将会夸大平均数。在这种情况下，中位数就是对中心位置更好的度量。

⊜ 每一年的增长因子等于1加上0.01乘以回报率。增长因子小于1，表明负增长，而增长因子大于1，表明正增长。增长因子不能小于0。

何平均数回答这个问题。由于 10 个增长因子的乘积为 1.334 493，几何平均数是 1.334 493 的 10 次方根[⊖]或

$$\bar{x}_g = \sqrt[10]{1.334\,493} = 1.029\,275$$

几何平均数告诉我们年收益平均以（1.029 275 - 1）×100% 或 2.927 5% 的速率增长。换句话说，对于 2.927 5% 的年平均增长率，第 1 年年初在该基金的 100 美元投资额，第 10 年年末将增加到 $100 \times 1.029\,275^{10} = 133.449\,3$ 美元。

了解年回报率的算术平均数不能提供这个投资的平均年增长率是非常重要的。表 3-2 中的 10 个年增长率之和为 50.4，因此，10 个年回报率的算术平均数为 50.4/10 = 5.04%。一个经纪人通过声称平均年回报率为 5.04%，试图说服你投资这个基金。这个声明不仅容易让人误导，还是错误的。平均年回报率 5.04% 对应的平均增长因子为 1.050 4。因此，如果平均增长因子确实为 1.050 4 的话，第 1 年年初在该基金投资 100 美元，第 10 年年末将增加到 $100 \times 1.050\,4^{10} = 163.451$ 美元。但是，利用表 3-2 中的年回报率，我们看到最初的 100 美元投资额第 10 年年末的价值为 133.45 美元。这个经纪人的平均年回报率为 5.04% 的声明极度夸大了该共同基金的实际增长率。这个问题是样本平均数只适合于加法过程。对于乘法过程，诸如增长率的应用，几何平均数是合适的位置度量。

在财务、投资和银行业的问题中，几何平均数的应用尤为常见，当你任何时候想确定过去几个连续时期的平均变化率时，都能应用几何平均数。其他通常的应用包括物种总体、农作物产量、污染水平以及出生率和死亡率的变化。注意，几何平均数也可以用于发生在所有时间长度的连续时期的任何数量的变化率。除了年变化率外，几何平均数也常常用于发现季度、月、周以及天的平均变化率。

3.1.5　众数

另一种位置度量是**众数**（mode）。众数的定义如下。

> **众数**
>
> 众数是出现次数最多的数据。

为了说明如何确定众数，我们考虑 5 个班级人数的样本。出现次数超过一次的数值只有 46。由于这个数值的出现次数为 2，是出现次数最多的数值，所以它就是众数。作为另外一个例子，考虑商学院毕业生起薪数据的样本。只有 3 880 美元是唯一出现次数超过一次的起始月薪数值。既然它出现的频数最多，它就是众数。

有时出现次数最多的数值可能有两个或更多。在这种情况下，存在不止一个众数。如果在数据中正好有两个众数，我们称数据集是双众数的。如果在数据中众数超过两个，我们称数据集是多众数的。在多众数的情况下，几乎从不报告众数。因为列出三个或以上的众数，对于描述数据的位置并不能起多大作用。

3.1.6　百分位数

百分位数（percentile）提供了数据如何散布在从最小值与最大值的区间上的信息。对于没有多个重复数值的数据，第 p 百分位数将数据分割为两个部分：大约有 p% 的观测值比第 p 百分位数小；而大约由（$100-p$）% 的观测值比第 p 百分位数大。第 p 百分位数的严格定义如下。

> **百分位数**
>
> 第 p 百分位数是满足下列条件的一个数值：至少有 p% 的观测值小于或等于该值，且至少有（$100-p$）% 的观测值大于或等于该值。

高等院校经常以百分位数的形式报告入学考试的成绩。比如，假设一名申请者入学考试的语言部分得到的原始分是 54 分。相对于参加同样考试的其他学生，这个学生的表现如何，人们可能看不清楚。但是，如果原始的 54 分恰好对应着第 70 百分位数，那么我们就能知道：大约有 70% 的学生的成绩比他低，大约有 30% 的学生成绩比他高。

⊖　使用大多数计算器或使用 Excel 中的 POWER 函数，可以计算 n 次方根。例如，使用 Excel，1.334 493 的 10 次方根 = POWER（1.334 493，1/10）或 1.029 275。

下面是用来计算第 p 百分位数的步骤[一]：

计算第 p 百分位数

第 1 步：把数据按升序排列（从小到大的顺序）。

第 2 步：计算指数 i

$$i = \left(\frac{p}{100}\right)n$$

式中，p 为所求的百分位数；n 为观测值的个数。

第 3 步：(a) 若 i 不是整数，则向上取整。大于 i 的下一个整数表示第 p 百分位数的位置。

(b) 若 i 是整数，则第 p 百分位数是第 i 项和第 $(i+1)$ 项数据的平均值。

为了举例说明上面的计算步骤，我们来确定表 3-1 中起薪数据的第 85 百分位数。

第 1 步：将数据按升序排列：

3 710　3 725　3 850　3 880　3 880　3 890　3 920　3 940　3 950　4 050　4 130　4 325

第 2 步：

$$i = \left(\frac{p}{100}\right)n = \left(\frac{85}{100}\right)12 = 10.2$$

第 3 步：由于 i 不是整数，故向上取整。第 85 百分位数的位置是大于 10.2 的下一个整数，即第 11 项。

再回到数据中，我们看出第 85 百分位数是第 11 项的数值，即 4 130。

作为另一个例子，我们考虑第 50 百分位数的计算。应用第 2 步，我们得到：

$$i = \left(\frac{50}{100}\right)12 = 6$$

由于 i 是整数，由第 3 步（b）可知，第 50 百分位数是第 6 项与第 7 项数据的平均值。因此第 50 百分位数是 $(3\,890 + 3\,920)/2 = 3\,905$。注意第 50 百分位数同时也是中位数。

3.1.7 四分位数

人们经常需要将数据划分为四部分，每一部分大约包含 1/4 或 25% 的观测值。这些分割点称为**四分位数**（quartiles）[二]，其定义是

Q_1 = 第一四分位数，或第 25 百分位数

Q_2 = 第二四分位数，或第 50 百分位数（也是中位数）

Q_3 = 第三四分位数，或第 75 百分位数

为了计算起薪数据的四分位数，我们将数据按升序排列。

3 710　3 725　3 850　3 880　3 880　3 890　3 920　3 940　3 950　4 050　4 130　4 325

我们已经确定 Q_2，即第二四分位数（中位数）为 3 905。计算 Q_1 和 Q_3，需要利用计算第 25 百分位数和第 75 百分位数的方法。这些计算如下。

对于 Q_1

$$i = \left(\frac{p}{100}\right)n = \left(\frac{25}{100}\right)12 = 3$$

由于 i 是整数，由计算法则的第 3 步（b）可知，第一四分位数即第 25 百分位数，是第 3 项与第 4 项数值的平均值。所以 $Q_1 = (3\,850 + 3\,880)/2 = 3\,865$。

对于 Q_3

[一] 遵循这些步骤将使得计算百分位数相当容易。

[二] 四分位数是一种特殊的百分位数，因此，计算百分位数的步骤可以直接用于四分位数的计算。

$$i=\left(\frac{p}{100}\right)n=\left(\frac{75}{100}\right)12=9$$

同样，由于 i 是整数，由计算法则的第3步（b）可知，第三四分位数即第75百分位数，是第9项与第10项数值的平均值。所以 $Q_3=(3\,950+4\,050)/2=4\,000$。

四分位数将起薪数据划分为四部分，每部分包含25%的观测值。

25%的数据			25%的数据			25%的数据			25%的数据		
3 710	3 725	3 850	3 880	3 880	3 890	3 920	3 940	3 950	4 050	4 130	4 325

$Q_1=3\,865$　　$Q_2=3\,905$（中位数）　　$Q_3=4\,000$

我们将四分位数分别定义为第25百分位数、第50百分位数和第75百分位数。因此，我们用和计算百分位数相同的方法计算四分位数。但是，有时也用其他方法来计算四分位数，而且根据所用方法的不同，计算出四分位数的实际值也会略有不同。不过，计算四分位数的所有方法的目的都是将数据划分为相等的四部分。

注释和评论

1. 当数据集中含有极端值时，使用中位数作为中心位置的度量比平均数更合适。有时，当存在极端值的情况下，我们使用调整平均数（trimmed mean）。从数据中删除一定比例最大值和最小值，然后计算剩余数据的平均值，就能得到调整平均数。例如，我们从数据中删除5%最小的数值和5%最大的数值，然后计算剩余数据的平均值，就得到5%调整平均数。对 $n=12$ 的起薪样本数据，$0.05\times12=0.6$。四舍五入得这个值为1，表明5%调整平均数将删除一个最小的数值和一个最大的数值。用10个剩余观测值计算的5%调整调整平均数为3 924.50。
2. 其他常用的百分位数是五分位数（第20百分位数、第40百分位数、第60百分位数和第80百分位数）和十分位数（第10百分位数、第20百分位数、第30百分位数、第40百分位数、第50百分位数、第60百分位数、第70百分位数、第80百分位数和第90百分位数）。

练习

方法

2. 考虑数据值10，20，21，17，16和12为一个样本，计算平均数和中位数。

4. 考虑下面的数据。

时期	回报率（%）	时期	回报率（%）
1	−6.0	4	2.0
2	−8.0	5	5.4
3	−4.0		

过去5个时期的平均增长率是多少？

☆6. 考虑数据值53、55、70、58、64、57、53、69、57、68和53的一个样本，计算平均数、中位数和众数。

应用

8. 在2007～2008年美国大学生篮球联赛（NCAA）赛季，男子篮球队试图创下3分投篮历史新高，场均19.07个（《美联社体育》，2009年1月24日）。为了试图阻止这么多的3分投篮，并且鼓励在禁区里面进攻，从2008～2009年篮球赛季开始，NCAA规则委员会将3分线从19英尺9英寸向后移到20英尺9英寸。由2008～2009年赛季NCAA的19场篮球比赛组成一个样本，下表是3分球投篮次数和3分球投篮命中次数的样本数据。

3分球投篮次数	3分球命中次数	3分球投篮次数	3分球命中次数
23	4	17	7
20	6	19	10
17	5	22	7
18	8	25	11
13	4	15	6
16	4	10	5
8	5	11	3
19	8	25	8
28	5	23	7
21	7		

a. 每场比赛3分求投篮的平均次数是多少？

b. 每场比赛3分球命中的平均次数是多少？

c. 较近的3分线，球员的命中率为35.2%。对新的3分线，球员的命中率是多少？

d. 在2008~2009年赛季，NCAA改变规则，将3分线向后移到20英尺9英寸的影响是什么？美联社体育（Associated Press Sport）的文章声称："3分线后移并没有使比赛发生显著的变化。"你是否赞同这个观点？请解释。

☆10.《华尔街日报》提供了消费者购买诸如独户住宅、汽油、互联网服务、报税服务和住院治疗等方面的费用（2007年1月2日）。由H&R Block提供服务的纳税申报准备费用的典型样本数据如下。

120 230 110 115 160 130 150 105 195 155
105 360 120 120 140 100 115 180 235 255

a. 计算平均数、中位数和众数。

b. 计算第一和第三四分位数。

c. 计算并解释第90百分位数。

12. 由约翰尼·盖尔克奇、吉姆·帕森斯和凯莉·库柯主演的系列情景喜剧《生活大爆炸》，是收视最高的电视节目之一。2011年9月22日首播了2011~2012播出季的前2集；第1集创下了14 100万的收视人数，第2集创下了14 700万的收视人数。2011~2012播出季的前21集的收视人数（单位：100万）如下表所示。

播出日期	收视人数（100万）	播出日期	收视人数（100万）
2011.9.22	14.1	2012.1.12	16.1
2011.9.22	14.7	2012.1.19	15.8
2011.9.29	14.6	2012.1.26	16.1
2011.10.6	13.6	2012.2.2	16.5
2011.10.13	13.6	2012.2.9	16.2
2011.10.20	14.9	2012.2.16	15.7
2011.10.27	14.5	2012.2.27	16.2
2011.11.3	16.0	2012.3.8	15.0
2011.11.10	15.9	2012.3.29	14.0
2011.11.17	15.1	2012.4.5	13.1
2011.12.8	14.0		

a. 计算收视人数的最大值和最小值。

b. 计算平均数、中位数和众数。

c. 计算第一和第三四分位数。

d. 2011~2012播出季电视观众是上升了还是下降了？请讨论。

14. 存在名为Unemp的文件中的数据给出了美国各州和哥伦比亚特区的2011年3月和2012年3月的失业率（Bureau of Labor Statistics website，2012年4月20日）。为了比较2011年3月和2012年3月的失业率，计算2011年3月和2012年3月的失业数据的第一四分位数、中位数和第三四分位数。这些统计量对全国失业率的变化的建议是什么？

☆16. 大学生的平均等级分是根据计算加权平均数得到的。在大多数学校中，用下列数值来分别代表各个等级：A（4），B（3），C（2），D（1），F（0）。州立大学的某个学生在结束了60个学分的课程学习之后，有9个学分获得A，15个学分获得B，33个学分获得C，3个学分获得D。

a. 计算这个学生的平均等级分。

b. 州立大学的学生们在前60个学分的课程学习中，应至少达到2.5个平均等级分才能被商学院录取。问该学生能被录取吗？

18. 根据425个工商管理硕士项目的调查，《美国新闻与世界报道》将印第安纳大学Kelley商学院排在全美国最佳工商管理项目的第20位（*America's Best Graduate School*，2009年）。排序的依据是对商学院院长和企业招聘人员的调查。要求每一名调查对象对学院硕士项目的整体学术质量评定分数，从1（勉强够格）到5（杰出）。利用下表中商学院院长和企业招聘人员评定的样本数据，计算加权平均得分。请讨论。

质量评估	商学院院长	企业招聘人员
5	44	31
4	66	34
3	60	43
2	10	12
1	0	0

20. 2004年年初假设你在Stivers共同基金中投资10 000美元，在Trppi共同基金中投资5 000美元。随后一年年末每项投资的价值如下表所示。哪只共同基金表现得更好？

年	Stivers 共同基金	Trppi 共同基金
2004	11 000	5 600
2005	12 000	6 300
2006	13 000	6 900
2007	14 000	7 600
2008	15 000	8 500
2009	16 000	9 200
2010	17 000	9 900
2011	18 000	10 600

22. 公司当前的价值是0.25亿美元。如果6年前公司的价值是0.1亿美元，则过去6年间公司的年平均增长率是多少？

3.2 变异程度的度量

除了位置的度量以外，人们往往还需要考虑变异程度亦即离散程度的度量。例如，假设你是一家大型制造公司的采购代理商，你经常向两个不同的供应商下订单。经过几个月的运营，你发现两个供货商完成订单所需的平均时间都是10天左右。二者完成订单所需时间的直方图如图3-2所示。尽管两个供货商的平均交货时间都是10天，但他们在按时交货方面是否拥有相同的可信度？注意直方图中交货时间的变异程度或离散程度，你会选择哪一家供货商呢？⊖

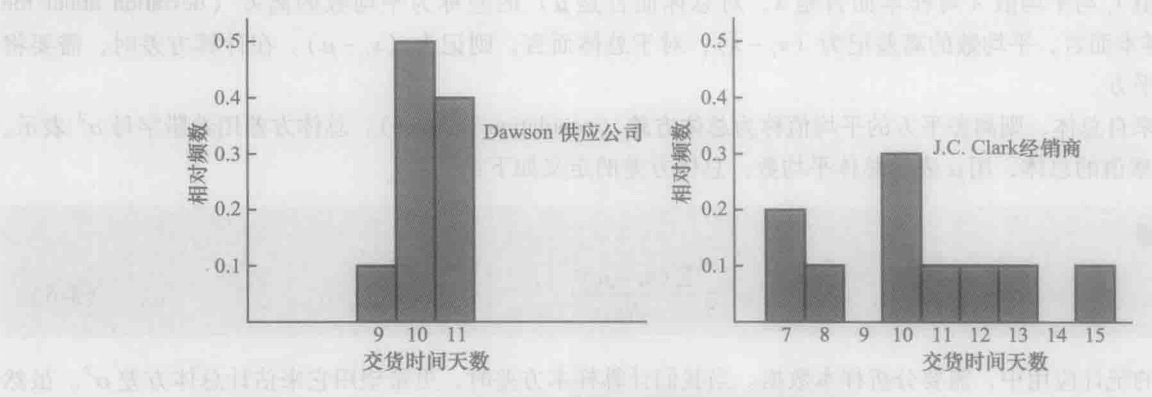


图3-2　完成订单所需时间的直方图

对于大多数公司来说，按时收到原材料和货物供给是很重要的。从J. C. Clark经销商的直方图来看，7天或8天交货可能会受到称赞，但是一部分延迟到13～15天的交货，对于协调公司员工的安排和按时完成生产认为将非常困难。这个例子说明了这样一种情形，即交货时间的变异程度往往是选择供应商的最主要考虑因素。对大多数的采购代理商来说，像Dawson供应公司这样变异程度较小的供应商，才是较理想的供应商。

现在我们来讨论一些常用的变异程度的度量。

3.2.1 极差

极差（range）是一种最简单的变异程度的度量。

极差

极差 = 最大值 - 最小值

让我们参阅表3-1中商学院毕业生的起薪数据。最高起薪为4 325美元，最低起薪为3 710美元。因此，极差为4 325 - 3 710 = 615（美元）。

尽管极差是最容易计算的变异程度的度量，但它很少被单独用来度量变异程度。原因是极差仅仅以两个观测值为依据，因此极易受到异常值的影响。假设某个毕业生的起薪为每月10 000美元，在这种情况下，极差将为10 000 - 3 710 = 6 290（美元），而不是615美元。如此之大的极差将不能准确地描述数据的变异程度，因为在12个起薪数据中有11个都集中在3 710～4 130。

3.2.2 四分位数间距

四分位数间距（interquartile range，IQR）作为变异程度的一种度量，能够克服异常值的影响。它是第三四分位数 Q_3 与第一四分位数 Q_1 的差值。也就是说，四分位数间距是在中间的50%的数据的极差。

⊖ 交货时间的变异性造成按时完成生产的不确定性，本节的方法有助于测量和了解变异性。

四分位数间距

$$IQR = Q_3 - Q_1 \tag{3-5}$$

对于起薪数据，四分位数 $Q_3 = 4\ 000$ 和 $Q_1 = 3\ 865$。因此，四分位数间距等于 4 000 − 3 865 = 135（美元）。

3.2.3 方差

方差（variance）是用所有数据对变异程度所做的一种度量。方差依赖于每个观察值（x_i）与平均值之间的差异，每个观察值 x_i 与平均值（对样本而言是 $\bar{x}$，对总体而言是 μ）的差称为平均数的离差（deviation about the mean）。对于样本而言，平均数的离差记为（$x_i - \bar{x}$）；对于总体而言，则记为（$x_i - \mu$）。在计算方差时，需要将平均数的离差平方。

如果数据来自总体，则离差平方的平均值称为**总体方差**（population variance），总体方差用希腊字母 σ^2 表示。对于有 N 个观察值的总体，用 μ 表示总体平均数，总体方差的定义如下：

总体方差

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \tag{3-6}$$

在大多数的统计应用中，需要分析样本数据。当我们计算样本方差时，更希望用它来估计总体方差 σ^2。虽然关于样本方差的详细解释已超出了本书的范围，但可以证明，如果样本平均数的离差平方和除以 $n-1$，而不是 n 时，所得到的样本方差㊀是总体方差的无偏估计。鉴于此，用 s^2 表示的样本方差（sample variance）定义如下：

样本方差

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{3-7}$$

为了举例说明样本方差的计算过程，我们使用第 3.1 节的 5 个大学班级样本的班级人数数据。在表 3-3 中给出了数据汇总，包括平均数的离差和平方离差的计算过程。平均数的离差平方和为 $\sum (x_i - \bar{x})^2 = 256$，并且 $n-1=4$。所以样本方差为

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{256}{4} = 64$$

表 3-3 班级人数数据的平均数的离差和平方离差的计算

班级学生数（x_i）	班级人数平均数（$\bar{x}$）	平均数的离差（$x_i - \bar{x}$）	平均数离差的平方（$x_i - \bar{x}$）^{2}
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		0	256
		$\sum (x_i - \bar{x})$	$\sum (x_i - \bar{x})^2$

我们注意到，样本方差的单位往往会引起人们的混淆。因为在方差的计算中，求和的数值 $(x_i - \bar{x})^2$ 都是平方，所以样本方差的单位也是平方。例如，班级人数数据的样本方差为 $s^2 = 64$（人）^{2}。方差的平方单位使得人们对于方差的数值很难找到直观的理解和诠释。我们建议您把方差作为比较两个或以上变量变异程度的有用工具。㊁在变

㊀ 样本方差 s^2 是总体方差 σ^2 的点估计。

㊁ 方差在比较两个或以上变量的变异程度时很有用。

量的比较中，拥有较大方差的变量显示其变异程度也较大，而对于方差数值的进一步解释可能就没有必要了。

作为计算样本方差的另一个例子，考虑表3-1中12名商学院毕业生的起薪数据。在3.1节我们计算出起薪的样本平均数为3 940。样本方差（$s^2=27\ 440.91$）的计算过程如表3-4所示。

表3-4 起薪数据的样本方差的计算

起始月薪（x_i）	样本平均数（$\bar{x}$）	平均数的离差（$x_i-\bar{x}$）	平均数离差的平方（$x_i-\bar{x}$）^{2}
3 850	3 940	−90	8 100
3 950	3 940	10	100
4 050	3 940	110	12 100
3 880	3 940	−60	3 600
3 755	3 940	−185	34 225
3 710	3 940	−230	52 900
3 890	3 940	−50	2 500
4 130	3 940	190	36 100
3 940	3 940	0	0
4 325	3 940	385	148 225
3 920	3 940	−20	400
3 880	3 940	−60	3 600
		0	301 850
		$\sum(x_i-\bar{x})$	$\sum(x_i-\bar{x})^2$

由式（3-7）

$$s^2=\frac{\sum(x_i-\bar{x})^2}{n-1}=\frac{301\ 850}{11}=27\ 440.91$$

在表3-3和表3-4中，我们列出了平均数的离差之和，也列出了平均数的平方离差之和。对任何数据集，平均数的离差之和将总是等于零。在表3-3和表3-4中，我们注意到$\sum(x_i-\bar{x})=0$。由于正的离差和负的离差总是相互抵消，从而导致平均数的离差之和等于零。

3.2.4 标准差

我们定义**标准差**（standard deviation）为方差的正平方根。沿用前面的样本方差和总体方差的记号，我们以s表示样本标准差[㊀]，以σ表示总体标准差。标准差可以用以下方法从方差推导而来。

标准差

$$样本标准差\ s=\sqrt{s^2} \tag{3-8}$$

$$总体标准差\ \sigma=\sqrt{\sigma^2} \tag{3-9}$$

我们记得，5个大学班级人数的样本方差$s^2=64$，因此样本标准差为$s=\sqrt{64}=8$。对于起薪数据，样本标准差为$s=\sqrt{27\ 440.91}=165.65$。

将方差转为与其相对应的标准差[㊁]有什么好处？我们知道方差的单位都是平方项，例如，对于商学院毕业生的起薪数据，样本方差为$s^2=27\ 440.91$（美元）^{2}。由于标准差是方差的平方根，因此方差的单位美元平方就转化为标准差的单位美元。所以，起薪数据的标准差为165.65美元。换句话说，标准差和原始数据的单位度量相同。标准差更容易和单位与平均数等其他统计量进行比较，这是由于它们的单位与原始数据是相同的。

㊀ 样本标准差s是总体标准差σ的点估计。

㊁ 标准差比方差更易于解释，因为标准差与数据的单位相同。

3.2.5 标准差系数

在某些情况下，我们可能对表示标准差相对于平均数大小的描述统计量感兴趣。这一量度称为**标准差系数**（coefficient of variation）[⊖]，它通常表示为百分数。

标准差系数

$$\left(\frac{\text{标准差}}{\text{平均数}}\times 100\right)\% \tag{3-10}$$

对于班级人数数据，我们知道样本平均数为44，样本标准差为8，因此标准差系数为［（8/44）×100］% = 18.2%。也就是说，标准差系数告诉我们，样本标准差是样本平均数的18.2%。对于起薪数据，样本平均数为3 940，样本标准差为165.65，因此标准差系数为［（165.65/3 940）×100］% = 4.2%，它说明样本标准差仅为样本平均数的4.2%。一般地，在比较具有不同标准差和不同平均数的变量的变异程度时，标准差系数是一个很有用的统计量。

注释和评论

1. 统计软件包和电子表格能用来计算本章所讲述的描述统计量。将数据输入工作表以后，通过使用一些简单的命令就能够产生所需要的结果。在本章末的三个附录中，我们将演示如何利用 Minitab、Excel 和 StatTools 来计算描述统计量。
2. 标准差通常用来度量与股票和股票基金投资相关的风险（Morningstar website，2012 年 7 月 21 日）。它给出了月回报率如何围绕长期平均回报率波动状况的度量。
3. 当使用计算器计算方差和标准差时，对样本平均数 $\bar{x}$ 和平方离差 $(x_i-\bar{x})^2$ 进行四舍五入处理，可能会引起方差或标准差的计算误差。为了减少舍入误差，我们建议在计算的中间步骤至少保留 6 位有效数字，而所得的方差或标准差可以保留少一些的位数。
4. 计算样本方差的另外一个公式为

$$s^2=\frac{\sum x_i^2-n\bar{x}^2}{n-1}$$

式中，$\sum x_i^2=x_1^2+x_2^2+\cdots+x_n^2$。

5. 平均绝对偏差（MAE）是另一种变异程度的度量，将每个观察值与平均值之间的差异绝对值求和再除以观测值的个数计算得出。对于一个容量为 n 的样本，MAE 的计算公式如下：

$$\text{MAE}=\frac{\sum|x_i-\bar{x}|}{n}$$

对于第 3.1 节的班级人数数据，$\bar{x}=44$，$\sum|x_i-\bar{x}|=28$，因此 MAE = 28/5 = 5.6。在第 17 章，你可以进一步学习 MAE 和其他变异程度的度量。

练习

方法

24. 考虑数据值为 10，20，12，17 和 16 的样本，计算方差和标准差。

应用

☆26. 一名保龄球手在六局比赛中的得分为 182，168，184，190，170 和 174。用这些数据作为一个样本，计算下列描述统计量：

a. 极差。 b. 方差。

c. 标准差。 d. 标准差系数。

28. 澳大利亚公开赛是每年第一个举行的职业网球四大满贯赛事。维多利亚·阿扎伦卡击败玛利亚·莎拉

⊖ 标准差系数是对变异程度的相对度量，它衡量标准差相对于平均数的大小。

波娃赢得了2012年澳大利亚公开赛女子单打冠军（《华盛顿邮报》，2012年1月27日）。在锦标赛期间，阿扎伦卡的发球速度达到每小时178公里。下面列出了2012年澳大利亚公开赛发球速度前20名的女子单打选手。

选手	发球速度（公里/小时）	选手	发球速度（公里/小时）
S. 威廉姆斯	191	G. 阿恩	179
S. 利斯基	190	V. 阿扎伦卡	178
M. 凯瑟	187	A. 伊万诺维奇	178
L. 夏迪嘉	187	P. 科维托娃	178
J. 加多索娃	187	M. 克拉吉塞克	178
J. 汉普顿	181	V. 杜舍维纳	178
B. 马泰克	181	S. 斯托瑟	178
F. 斯齐亚沃尼	179	S. 克里斯蒂	177
P. 帕门蒂尔	179	M. 巴瑟尔	177
N. 佩特洛娃	179	P. Ormaechea	177

a. 计算发球速度的平均数、方差和标准差。

b. 由2011年温布尔登锦标赛发球速度前20名的女子单打选手组成一个类似的样本，得到发球速度的样本平均数为每小时182.5公里，方差和标准差分布为33.3和5.77。讨论澳大利亚公开赛和温布尔登女子锦标赛发球速度之间的差异。

30. 下面的数据用于绘制直方图，这些数据分别是Dawson供应公司和J. C. Clark经销商完成订单所需的时间（见图3-2）。

Dawson供应公司的交货时间：

11　10　9　10　11　11　10　11　10　10

Clark经销商的交货时间：

8　10　13　7　10　11　10　7　15　12

用极差和标准差来证明前面的观察结果：Dawson供应公司的交货时间更一致和更可靠。

32. 美国零售业联盟报告大学一年级学生比其他年级的大学生在上学项目上花费更多（《今日美国》，2006年8月4日）。25名一年级学生和20名四年级学生在上学上花费的样本数据存在名为BackToSchool的文件中。

a. 每个年级在上学上花费的均值是多少？这是否与美国零售业联盟的报告一致？

b. 每个年级花费的极差是多少？

c. 每个年级花费的四分位数间距是多少？

d. 每个年级花费的标准差是多少？

e. 在花费上一年级学生和四年级学生哪个变异更大？

34. 下面的数据是一所大学田径队的队员跑0.25英里和1英里的时间（单位：分钟）。

0.25英里的时间：0.92　0.98　1.04　0.90　0.99

1英里的时间：4.52　4.35　4.60　4.70　4.50

一位教练看到这个的样本后认为，跑0.25英里的时间更具有一致性。用标准差和标准差系数来汇总数据的变异性。标准差系数是否说明教练的说法成立。

3.3 分布形态、相对位置的度量以及异常值的检测

我们已经讲述了几种对数据位置和变异程度的度量方法，但对分布形态的度量往往也是重要的。在第2章中，我们注意到，直方图对分布的形态提供了一种很好的图形描述。分布形态的一种重要的数值度量被称为**偏度**（skewness）。

3.3.1 分布形态

图3-3是根据相对频数分布绘制的4个直方图。图3-3a和图3-3b中的直方图呈现一定程度的偏态：图3-3a的直方图是左偏，它的偏度为−0.85；图3-3b的直方图是右偏，它的偏度为+0.85。图3-3c的直方图是对称的，它的偏度为0。图3-3d的直方图严重右偏，它的偏度为1.62。用来计算偏度的公式有些复杂[⊖]，但是使用统计软件很容易计算偏度。对于左偏的数据，偏度是负数；对于右偏的数据，偏度是正值。如果数据是对称的，则偏度为0。

对于一个对称的分布，平均数和中位数是相等的。当数据的偏度是正值时，通常平均数比中位数要大；当数据的偏度是负值时，通常平均数比中位数要小。用于绘制图3-3d中直方图的数据是消费者在一家妇女服饰店购物的金额。平均购物金额是77.60美元，中位数是59.70美元。少数较大的购物金额将平均数拉大，但中位数不受

⊖ 计算样本数据偏度的公式为：偏度 $=\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3$

影响。当数据严重偏离时，中位数是位置的首选度量。

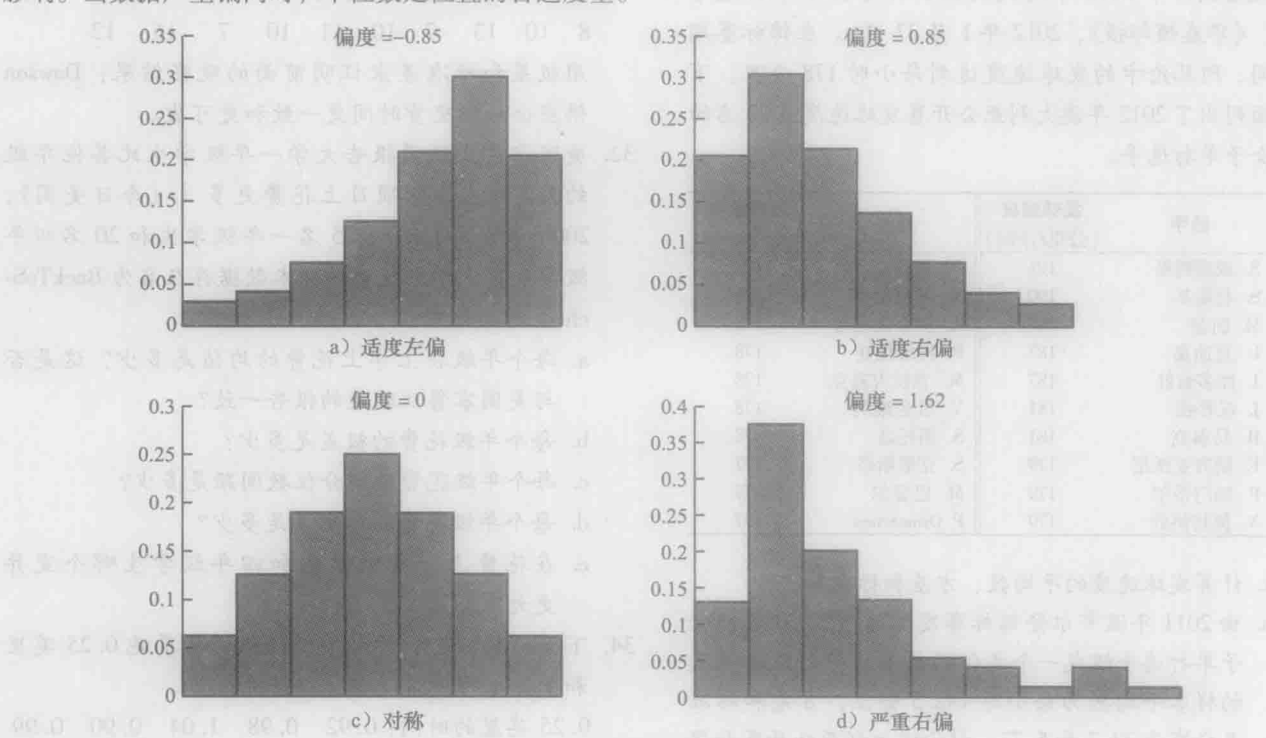


图 3-3 四个呈现不同偏度的分布的直方图

3.3.2 z-分数

对一个数据集，除了位置、变异程度和形态的度量外，我们还对数据集中的数值的相对位置感兴趣。相对位置的度量值能帮助我们确定一个特殊的数值距平均数有多远。

利用平均数和标准差，我们可以确定任何观察值的相对位置。假设我们有一个 n 个观测值 $x_1, x_2, \cdots, x_n$ 的样本。并且假设样本平均数 $\bar{x}$ 和样本标准差 s 已经被计算出来。与任何一个数值 x_i 有关的另一个数值称为 x_i 的 **z-分数**（z-score）。式（3-11）是计算每个 x_i 的 z-分数的公式。

z-分数

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3-11}$$

式中，z_i 代表 x_i 的 z-分数；$\bar{x}$ 为样本平均数；s 为样本标准差。

z-分数往往被称为标准化数值。z-分数 z_i，能被解释为 x_i 与平均数 $\bar{x}$ 的距离是 z_i 个标准差。例如，$z_1 = 1.2$，表示 x_1 比样本平均数大 1.2 个标准差。类似地，$z_2 = -0.5$，表示 x_2 比样本平均数小 0.5 或 1/2 个标准差。当观察值大于平均数时，z-分数将大于零；当观察值小于平均数，z-分数将小于零。z-分数等于零，则表示观察值等于平均数。

任何观测值的 z-分数都被认为是对数据集中观测值相对位置的量度。因此，两个不同数据集的观测值具有相同的 z-分数，则可以说它们具有相同的相对位置，因为它们与平均数的距离有相同个数的标准差。

表 3-5 中计算了第 3.1 节中班级人数数据的 z-分数[⊖]。我们已知样本平均数 $\bar{x} = 44$，样本标准差 $s = 8$。第 5 个观测值的 z-分数为 −1.50，说明它距离平均数最远，且比平均数小 1.50 个标准差。图 3-4 给出了班级人数数据的

⊖ 一个变量的数值转换成 z-分数的过程常常被称为 z 变换。

打点图，在打点图坐标轴下方有相对应的 z- 分数的图形表示。

表 3-5 班级人数数据的 z- 分数

班级人数（x_i）	平均数的离差（$x_i-\bar{x}$）	z- 分数$\left(\frac{x_i-\bar{x}}{s}\right)$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	−2	−2/8 = −0.25
46	2	2/8 = 0.25
32	−12	−12/8 = −1.50

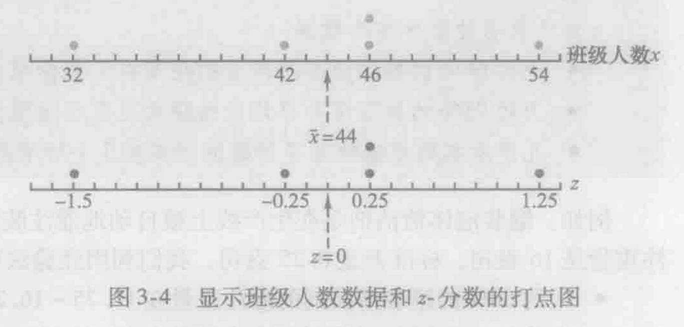


图 3-4 显示班级人数数据和 z- 分数的打点图

3.3.3 切比雪夫定理

切比雪夫定理（Chebyshev's theorem）能使我们指出与平均数的距离在某个特定个数的标准差之内的数据值所占的比例。

切比雪夫定理

与平均数的距离在 z 个标准差之内的数据项所占比例至少为 $(1-1/z^2)$，其中 z 是大于 1 的任意实数[1]。

当 $z=2$，3 和 4 个标准差时，该定理的一些应用如下：

- 至少 0.75 或 75% 的数据值与平均数的距离在 $z=2$ 个标准差之内。
- 至少 0.89 或 89% 的数据值与平均数的距离在 $z=3$ 个标准差之内。
- 至少 0.94 或 94% 的数据值与平均数的距离在 $z=4$ 个标准差之内。

应用切比雪夫定理的一个实例是，假设某大学 100 名学生商务统计课期中考试的平均成绩为 70 分，标准差为 5 分。那么有多少学生的考试成绩在 60 ~ 80 分？有多少学生的考试成绩在 58 ~ 82 分？

对于 60 ~ 80 分的考试成绩，我们注意到 60 比平均数低 2 个标准差，而 80 比平均数高 2 个标准差。利用切比雪夫定理，我们知道至少有 0.75 或 75% 的观测值与平均数的距离在 2 个标准差之内。因此，至少有 75% 的学生考试成绩在 60 ~ 80 分。

对于 58 ~ 82 分的考试成绩，我们看到，$(58-70)/5=-2.4$，表明 58 比平均数低 2.4 个标准差；而 $(82-70)/5=+2.4$，表明 82 比平均数高 2.4 个标准差。应用切比雪夫定理且 $z=2.4$，我们得到：

$$\left(1-\frac{1}{z^2}\right)=\left(1-\frac{1}{2.4^2}\right)=0.826$$

即至少有 82.6% 的学生考试成绩在 58 ~ 82 分。

3.3.4 经验法则

切比雪夫定理的优点之一就是它适用于任何数据集而不论其数据分布的形状。的确，它可以用于图 3-3 中任何一个分布。但是，在实际应用中，人们发现许多数据集具有类似图 3-5 所示的对称的峰形或钟形分布。当数据被认为近似于这种分布时，就可以运用**经验法则**（empirical rule）来确定与平均数的距离在某个特定个数的标准差之内的数据值所占的比例。[2]

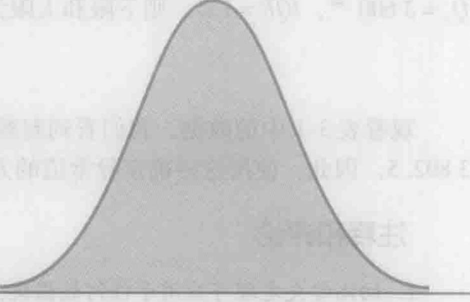


图 3-5 对称的峰形或钟形分布

[1] 切比雪夫定理要求 $z>1$，但 z 不一定是整数。

[2] 经验法则以正态分布为依据，该分布将在第 6 章中予以介绍。正态分布广泛应用在本书中。

经验法则

对于具有钟形分布的数据：

- 大约 68% 的数据值与平均数的距离在 1 个标准差之内。
- 大约 95% 的数据值与平均数的距离在 2 个标准差之内。
- 几乎所有的数据值与平均数的距离在 3 个标准差之内。

例如，罐装液体清洁剂是在生产线上被自动地灌注装满，灌注的重量通常具有钟形分布的特点。如果平均灌注重量是 16 盎司，标准差是 0.25 盎司，我们利用经验法可以得到下面的结论：

- 大约 68% 的罐装清洁剂的灌注重量在 15.75 ~ 16.25 盎司之间（即与平均数的距离在 1 个标准差之内）。
- 大约 95% 的罐装清洁剂的灌注重量在 15.50 ~ 16.50 盎司之间（即与平均数的距离在 2 个标准差之内）。
- 几乎所有的罐装清洁剂的灌注重量在 15.25 ~ 16.75 盎司之间（即与平均数的距离在 3 个标准差之内）。

3.3.5 异常值的检测

有时数据集中会包含一个或多个数值异常大或异常小的观测值，这样的极端值称为**异常值**（outliers）。有经验的统计人员会采取一定的步骤来识别异常值[㊀]，并对它们仔细鉴别。异常值可能是一个被错误记录的数据值，如果是这样，则可以在进一步分析之前把它更正。异常值也可能是一个被错误包含在数据集中的观测值，如果是这样，则可以把它删除。最后，异常值也可能就是一个反常的数据值，它被正确地记录并且属于数据集。这种情况下，它应该被保留。

标准化数值（z-分数）可以用来确认异常值。经验法则告诉我们，对于具有钟形分布的数据，几乎所有的数据值与平均数的距离都在 3 个标准差之内。因此，在利用 z-分数来检测异常值时，我们建议把 z-分数小于 -3 或大于 $+3$ 的任何数值都视为异常值。然后，对它们的准确性进行检查，以确定它们是否属于数据集。

参见表 3-4 中班级人数数据的 z-分数。第 5 项的 z-分数为 -1.50，表明它距离平均值最远。但是，这个标准化数值仍在判定异常值的界限 -3 到 $+3$ 之内。因此，z-分数不能表明，在班级人数数据中存在异常值。

确定异常值的另一种方法是以第一四分位数和第三四分位数（Q_1 和 Q_3）以及四分位数间距（IQR）为依据。利用这种方法，我们首先计算如下的下线和上限：

$$\text{下限} = Q_1 - 1.5 \times IQR$$

$$\text{上限} = Q_3 + 1.5 \times IQR$$

如果一个观测值的数值小于下限或大于上限，则被归类为异常值[㊁]。对于表 3-1 的起始月薪数据，$Q_1 = 3\,465$，$Q_3 = 3\,600$[㊂]，$IQR = 135$。则下限和上限为

$$3\,465 - 1.5 \times 135 = 3\,262.5$$

$$3\,600 + 1.5 \times 135 = 3\,802.5$$

观看表 3-1 中的数据，我们看到起薪数据的观测值没有小于下限 3 262.5。但是，有一个起薪 3 925 大于上限 3 802.5。因此，使用这种确定异常值的方法，3 925 被认定为异常值。

注释和评论

1. 切比雪夫定理可应用于任何数据集，并能够估计与平均数在一定个数的标准差之内的最少数量的数据项。如果已知数据集近似呈现钟形分布，则有更多的数据项在此范围之内。例如，经验法则告诉我们，大约有 95% 的数据值与平均数的距离在 2 个标准差之内；而切比雪夫定理仅仅告诉我们，至少有 75% 的数据项在此范围之内。

㊀ 在以数据分析为依据作出决策之前，检测异常值是有意义的。在记录数据和将数据输入计算机时，经常会有误差。异常值并非都要删除，但它们的精确度和适应性应予以核实。

㊁ 用第一和第三四分位数以及四分位数间距确定异常值的方法，不必与依据 z-分数小于 -3 或大于 3 的方法相同。可选用一种或两种方法。

㊂ 原书上此处与下面的下限和上限都有误，翻译版已更正。

2. 在对数据集进行分析之前，统计人员经常会对数据进行各种检查，以确保其有效性。在大型研究中，在记录数据或将数据输入计算机时往往会发生误差。检测异常值是检查数据有效性的一个工具。

练　习

方法

36. 考虑平均数为500，标准差为100的一个样本，数据值为520，650，500，450和280的z-分数是多少？

38. 假设数据具有钟形分布，其平均数为30，标准差为5。利用经验法则来确定在下列范围之内的数据所占的比例。

a. 20～40　b. 15～45　c. 25～35

应用

40. 美国能源信息署报告，每加仑普通汽油的平均零售价为3.43美元（Energy Information Administration，2012年7月）。假设每加仑汽油的零售价具有钟形分布，标准差为0.10美元。

a. 每加仑普通汽油的零售价在3.33～3.53美元的比例是多少？

b. 每加仑普通汽油的零售价在3.33～3.63美元的比例是多少？

c. 每加仑普通汽油的零售价超过3.63美元的比例是多少？

42. 在加利福尼亚州，许多家庭用后院小屋建造家庭办公室、艺术工作室和业余爱好的储藏室。单个木制的后院建筑物的平均价格为3 100美元。假设标准差为1 200美元。

a. 后院建筑物的成本为2 300美元的z-分数是多少？

b. 后院建筑物的成本为4 900美元的z-分数是多少？

c. 解释（a）和（b）中的z-分数，对它们是否都被认为是异常值进行评价。

d. 在加利福尼亚州的Albany建造后院小屋办公室的价格为13 000美元。这个建筑物是否被认为是异常值？为什么？

44. 由10支NCAA大学篮球队组成的一个样本，提供的数据如下。

获胜队	获胜队得分	失败队	失败队得分	得分差
Arizona	90	Oregon	66	24
Duke	85	Georgetown	66	19
Florida State	75	Wake Forest	70	5
Kansas	78	Colorado	57	21
Kentucky	71	Notre Dame	63	8
Louisville	65	Tennessee	62	3
Oklahoma State	72	Texas	66	6
Purdue	76	Michigan State	70	6
Stanford	77	Southern Cal	67	10
Wisconsin	76	Illinois	56	20

a. 计算获胜队得分的平均数和标准差。

b. 假设所有NCAA比赛中获胜队的得分具有钟形分布。用（a）计算的平均数和标准差，估计所有NCAA比赛中获胜队得分在84分及以上的比例；估计所有NCAA比赛中，获胜队得分超过90分的比例。

c. 计算得分差的平均数和标准差，得分差数据是否包含异常值？请解释。

3.4　五数概括法和箱形图

以汇总统计为依据的汇总统计量和易画的图形，可以用来快速汇总大量数据。本节，我们将演示如何使用五数概括法和箱形图来确定一个大的数据集的几个特征。

3.4.1　五数概括法

在五数概括法（five-number summary）中，使用下面五个数来汇总数据：

（1）最小值

（2）第一四分位数（Q_1）

（3）中位数（Q_2）

（4）第三四分位数（Q_3）

（5）最大值

实施五数概括法首先将数据按升序排列，然后确定最小值、三个四分位数和最大值。对表3-1中的12个商学院毕业生的起始月薪数据，再次按照升序排列如下：

3 710　3 755　3 850　|　3 880　3 880　3 890　|　3 920　3 940　3 950　|　4 050　4 130　4 325

$Q_1 = 3\,865$　　$Q_2 = 3\,905$（中位数）　　$Q_3 = 4\,000$

在3.1节中我们已经计算了中位数为3 905，四分位数 $Q_1 = 3\,865$ 和 $Q_3 = 4\,000$。观察上述数据，可知最小值为3 710，最大值为4 325。于是，起薪数据的五数概括为：3 710，3 865，3 905，4 000，4 325。大约有1/4或25%的观测值在五数概括法的相邻两个数之间。

3.4.2 箱形图

箱形图（box plot）是基于五数概括法的数据图形汇总。[⊖]绘制箱形图的关键是计算四分位数间距 $IQR = Q_3 - Q_1$。图3-6是起始月薪数据的箱形图。绘制箱形图的步骤如下。

（1）画一个箱体，其边界分别是第一四分位数和第三四分位数。对于起始月薪数据，$Q_1 = 3\,865$，$Q_3 = 4\,000$。这个箱体包含了中间的50%的数据。

（2）在箱体上中位数的位置画一条垂线（对起薪数据，中位数为3 905）。

（3）利用四分位数间距 $IQR = Q_3 - Q_1$，设定界限在 Q_1 左侧的1.5个 IQR 处和在 Q_3 右侧的1.5个 IQR 处。对起始月薪数据，$IQR = Q_3 - Q_1 = 4\,000 - 3\,865 = 135$。因此，界线分别为：$3\,865 - 1.5 \times 135 = 3\,662.5$ 和 $4\,000 + 1.5 \times 135 = 4\,202.5$。界线以外的数据被认为是异常值。

（4）图3-6中的横线称作**触须线**（whisker）。触须线从箱体的边界一直画到步骤3计算出的界线以内的原始数据的最大值和最小值处。触须线的两个端点是3 710和4 130薪金值。

（5）最后，每个异常值的位置用符号 * 来标出。在图3-6中，我们看到有一个异常值4 325。

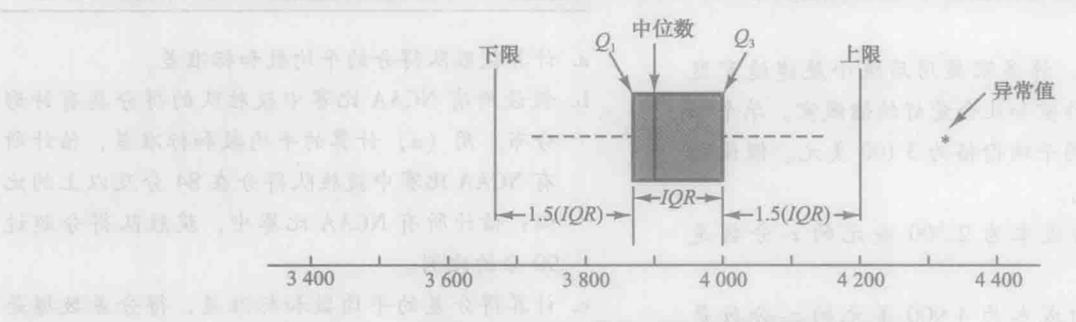


图3-6　显示下限和上限的起薪数据的箱形图

在图3-6中，我们包括表示下限和上限位置的线条。画这些线是用来说明如何计算界限值以及它们的位置所在。虽然总是需要计算界限，但通常情况下，箱形图中并不画出它们。图3-7是起薪数据箱形图的一般形状。

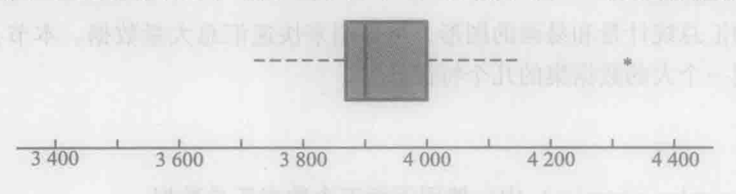


图3-7　月起薪数据的箱形图

为了比较商学院各专业毕业生的月起薪情况，选择了111名新近毕业的学生组成一个样本，记录每一名毕业

⊖ 箱型图提供了数据集几个特征的方便直观地显示。

生的专业和月起薪。图3-8是会计、金融、信息系统、管理和市场营销专业的 Minitab 箱形图。注意专业在横轴上，每一个箱形图出现在相对应专业的纵轴上。在对两个或多个组进行对比时，以这种方式展示箱形图是一个很好的图形技术。

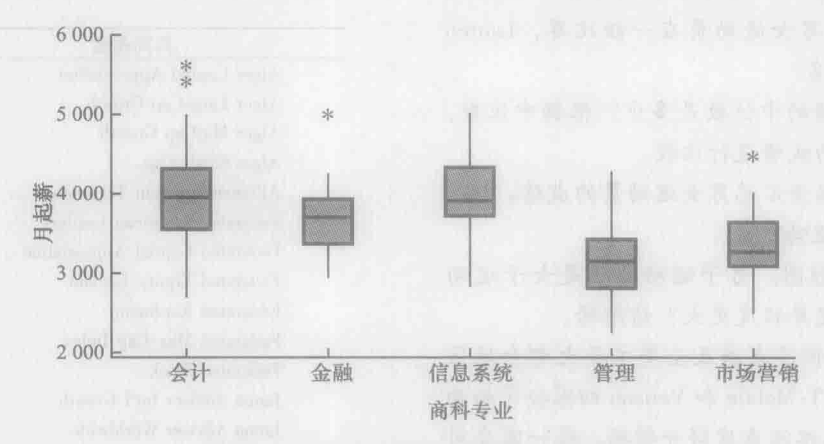


图3-8　各专业月起薪的 Minitab 箱形图

对于图3-8各专业月起薪的箱形图，你能得出怎样的观测结果？特别地，我们注意到：

- 会计专业的起薪较高；而管理和市场营销专业的起薪较低。
- 根据中位数，会计、信息系统有相似和较高的中位数，金融次之，而管理和市场营销显示有较低的中位数。
- 在会计、金融和市场营销专业中存在高起薪的异常值。
- 金融起薪的变异性最小，会计起薪的变异性最大。

基于这些箱形图，也许你可以看到更多的解释。

注释和评论

1. 探索性数据分析的一个优点就是使用简便，且只需进行少量的数学计算。我们只需要简单地将数据按升序排列，就可以确定五数概括法中的五个数值。然后，我们就能绘制箱形图，而不必计算数据的平均数和标准差。
2. 在附录3A，我们将演示如何使用 Minitab 来绘制起薪数据的箱形图。这样得到的箱形图看起来就像图3-7的那个箱形图，只是方向旋转了90°。

练习

方法

46. 考虑数据值为27，50，20，15，30，34，28和25的一个样本。用五数概括法来汇总数据。

☆48. 用五数概括汇总数据：5，15，18，10，8，12，16，10和6，并绘制箱形图。

应用

50. 每年1月在佛罗里达州那不勒斯举行半程马拉松比赛。比赛吸引全美以及世界各国的顶级运动员。2009年1月有22名男子和31名女子参加了19～24岁年龄组的比赛，比赛成绩（单位：分）如右所示（*Naples Daily*，2009年1月19日）。

名次	男子	女子	名次	男子	女子
1	65.30	109.03	17	128.40	132.03
2	66.27	111.22	18	130.90	133.20
3	66.52	111.65	19	131.80	133.50
4	66.85	111.93	20	138.63	136.57
5	70.87	114.38	21	143.83	136.75
6	87.18	118.33	22	148.70	138.20
7	96.45	121.25	23		139.00
8	98.52	122.08	24		147.18
9	100.52	122.48	25		147.35
10	108.18	122.62	26		147.50
11	109.05	123.88	27		147.75
12	110.23	125.78	28		153.88
13	112.90	129.52	29		154.83
14	113.52	129.87	30		189.27
15	120.95	130.72	31		189.28
16	127.98	131.67			

a. 佐治亚州Marietta的George Towett获得男子的第一名，佛罗里达州Gainesville的Lauren Wald获得女子的第一名。比较男子和女子第一名的成绩，如果53名男女运动员在一组比赛，Lauren Wald将是第几名？

b. 男女运动员成绩的中位数是多少？根据中位数，对男女运动员的成绩进行比较。

c. 分别用五数概括法汇总男女运动员的成绩。

d. 每个组有异常值吗？

e. 绘制两组的箱形图。男子运动员还是女子运动员比赛成绩的变异程度更大？请解释。

52.《消费者报告》提供了在遍及全美主要大都会地区对AT&T，Sprint，T-Mobile和Verizon四家公司移动电话服务的客户总体满意度得分数据。每一家公司的评分反映了客户对诸如费用、连线、掉线、静电干扰和客户支持等一系列因素的总体满意度。满意度得分从0~100，0表示完全不满意，100表示完全满意。20个大都会地区的四家移动电话服务公司的得分如下（《消费者报告》，2009年1月）：

大都会地区	AT&T	Sprint	T-Mobile	Verizon
亚特兰大	70	66	71	79
波士顿	69	64	74	76
芝加哥	71	65	70	77
达拉斯	75	65	74	78
丹佛	71	67	73	77
底特律	73	65	77	79
杰克逊维尔	73	64	75	81
拉斯维加斯	72	68	74	81
洛杉矶	66	65	68	78
迈阿密	68	69	73	80
明尼阿波利斯	68	66	75	77
费城	72	66	71	78
菲尼克斯	68	66	76	81
圣安东尼奥	75	65	75	80
圣迭戈	69	68	72	79
旧金山	66	69	73	75
西雅图	68	67	74	77
圣路易斯	74	66	74	79
坦帕	73	63	73	79
华盛顿	72	68	71	76

a. 首先考虑T-Mobile，得分中位数是多少？

b. 用五数概括法汇总T-Mobile的数据。

c. T-Mobile有异常值吗？请解释。

d. 对其他三家移动电话服务公司重复做（b）和（c）。

e. 在一张图上绘制出四家移动电话服务公司的箱形图。对四家公司的箱形图进行比较，讨论比较的结果。根据客户总体满意度，哪家是《消费者报告》推荐的最佳移动电话服务公司？

54. 下表是46个共同基金12个月的总回报率数据（*Smart Money*，2004年2月）。

共同基金	回报率（%）
Alger Capital Appreciation	23.5
Alger LargeCap Growth	22.8
Alger MidCap Growth	38.3
Alger Small Cap	41.3
AllianceBernstein Technology	40.6
Federated American Leaders	15.6
Federated Captial Appreciation	12.4
Federated Equity-Income	11.5
Federated Kaufmann	33.3
Federated Max-Cap Index	16.0
Federated Stock	16.9
Janus Adviser Int'l Growth	10.3
Janus Adviser Worldwide	3.4
Janus Enterprise	24.2
Janus High-Yield	12.1
Janus Mercury	20.6
Janus Overseas	11.9
Janus Worldwide	4.1
Nations Convertible Secuities	13.6
Nations Int'l Equity	10.7
Nations LargeCap Enhd. Core	13.2
Nations LargeCap Index	13.5
Nation MidCap Index	19.5
Nations Small Company	21.4
Nations SmallCap Index	24.5
Nations Strategic Growth	10.4
Nations Value Inv	10.8
One Group Diversified Equity	10.0
One Group Diversified Int'l	10.9
One Group Diversified Mid Cap	15.1
One Group Equity Income	6.6
One Group Int'l Equity Index	13.2
One Group Large Cap Growth	13.6
One Group Large Cap Value	12.8
One Group Mid Cap Growth	18.7
One Group Mid Cap Value	11.4
One Group Small Cap Growth	23.6
PBHG Growth	27.3
Putnam Europe Equity	20.4
Putnam Int'l Capital Opportunity	36.6
Putnam International Equity	21.5
Putnam Int'l New Opportunity	26.3
Strong Advisor Mid Cap Growth	23.7
Strong Growth 20	11.7
Strong Growth Inv	23.2
Strong Large Cap Growth	14.5

a. 这些共同基金回报率的平均数和中位数是多少？

b. 第一四分位数和第三四分位数各为多少？

c. 用五数概括法汇总数据。

d. 数据是否包含异常值？绘制箱形图。

3.5 两变量间关系的度量

迄今为止，我们已经介绍了在同一时间对一个变量数据汇总的数值方法。但是管理者或决策者常常更关心的是两个变量之间的关系。在本节我们将介绍描述两个变量间关系的度量：协方差和相关系数。

首先，我们重新考虑第2.4节中关于旧金山音像设备商店的例子。商店经理想确定周末电视广告播出次数与下周商店销售额之间的关系。表3-6中是销售额（单位：100美元）的样本数据，有10个观测值（$n=10$），每周一个观测值。图3-9的散点图中显示二者是正相关关系，较高的销售额（y）对应着较多的广告次数（x）。事实上，散点图提示我们，可以用一条直线来近似表达这种关系。在下面的讨论中，我们将介绍描述两变量间线性关系度量的**协方差**（covariance）。

表3-6 音像设备商店的样本数据

周次	电视广告数 x	销售额（100美元）y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

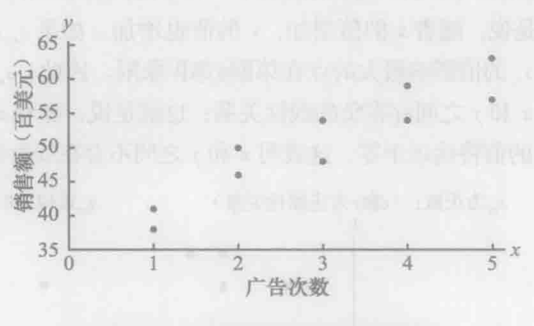


图3-9 音像设备商店的散点图

3.5.1 协方差

对于一个容量为 n 的样本，其观测值为（x_1，y_1），（x_2，y_2），…，（x_n，y_n），样本协方差的定义如下：

样本协方差

$$s_{xy}=\frac{\sum(x_i-\bar{x})\ (y_i-\bar{y})}{n-1} \tag{3-12}$$

在这个公式中，一个 x_i 与一个 y_i 相对应。我们将每个 x_i 与其样本平均数 $\bar{x}$ 的离差乘以对应的 y_i 与其样本平均数 $\bar{y}$ 的离差，再将所得的结果加总，然后除以 $n-1$。

在音像设备商店问题中，为了度量广告次数 x 与销售额 y 之间线性关系的强度，我们利用式（3-12）来计算样本协方差。$\sum(x_i-\bar{x})(y_i-\bar{y})$ 的计算过程如表3-7中所示。注意到 $\bar{x}=30/10=3$ 并且 $\bar{y}=510/10=51$。利用式（3-12），我们可以得出样本协方差为

$$s_{xy}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}=\frac{99}{9}=11$$

表3-7 样本协方差的计算

	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
合计	30	510	0	0	99

$$s_{xy}=\frac{\sum(x_i-\bar{x})\ (y_i-\bar{y})}{n-1}=\frac{99}{10-1}=11$$

计算容量为 N 的总体协方差的公式与式（3-12）相似。但是我们使用不同的符号，表示正在处理的是整个总体。

总体协方差

$$\sigma_{xy}=\frac{\sum(x_i-\mu_x)\ (y_i-\mu_y)}{N} \tag{3-13}$$

在式（3-13）中我们用符号 μ_x 表示变量 x 的总体平均数，用符号 μ_y 表示变量 y 的总体平均数，用 σ_{xy} 表示容量为 N 的总体协方差。

3.5.2 协方差的解释

为了有助于对样本协方差进行解释，如图3-10所示。它与图3-9的散点图相同，只是在 $\bar{x}=3$ 处有一条垂直虚线，在 $\bar{y}=51$ 处有一条水平虚线。这两条线将图划分为四个象限。第Ⅰ象限的点对应于 x_i 大于 $\bar{x}$ 且 y_i 大于 $\bar{y}$，第Ⅱ象限的点对应于 x_i 小于 $\bar{x}$ 且 y_i 大于 $\bar{y}$，依此类推。因此，$(x_i-\bar{x})(y_i-\bar{y})$ 的值，在第Ⅰ象限的点一定为正，第Ⅱ象限的点为负，在第Ⅲ象限的点为正，第Ⅳ象限的点为负。

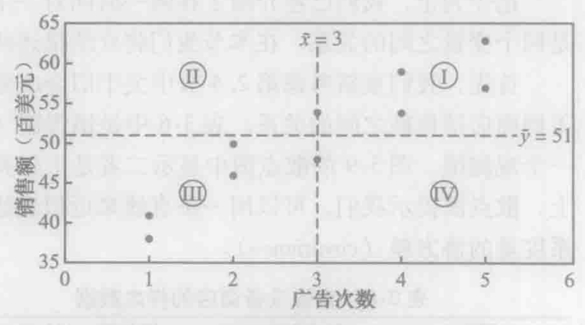


图3-10 音像设备商店的分区散点图

如果 s_{xy} 的值是正的，那么对 s_{xy} 的值影响最大的点必然在第Ⅰ和Ⅲ象限。因此，s_{xy} 为正值表示 x 和 y 之间存在正的线性关系[⊖]；也就是说，随着 x 的值增加，y 的值也增加。如果 s_{xy} 的值是负的，则对 s_{xy} 的值影响最大的点在第Ⅱ和第Ⅳ象限。因此，s_{xy} 为负值，则表示 x 和 y 之间存在负的线性关系；也就是说，随着 x 的值增加，y 的值减少。最后，如果各点在四个象限中分布均匀，则 s_{xy} 的值将接近于零，这表明 x 和 y 之间不存在线性关系。图3-11显示了与三种不同类型的散点图相对应的 s_{xy} 值。

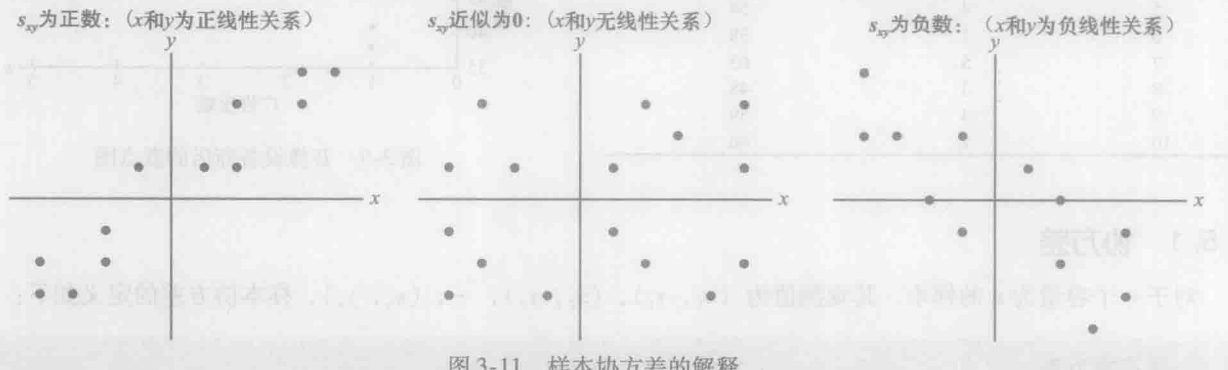


图3-11 样本协方差的解释

再次参见图3-10，我们可以看到音像设备商店的散点图与图3-11中顶部的图形相似。正像我们所期望的，样本协方差 $s_{xy}=11$ 表示正的线性关系。

从上面的讨论中可以看出，似乎协方差是一个大的正值就表示强的正线性相关关系，一个大的负数就表示强的负线性相关关系。但是，在使用协方差作为线性关系强度的度量时，一个问题就在于协方差的值依赖于 x 和 y 的计量单位。例如，假设我们要研究人的身高 x 与体重 y 的关系。显然，无论用英尺还是英寸来度量身高，二者关系的强度都应该相同。但是，如果用英寸来度量身高的话，我们计算出的 $(x_i-\bar{x})$ 的数值将比用英尺度量时要大得多。因此，用英寸来度量身高时，式（3-12）中的分子 $\sum(x_i-\bar{x})(y_i-\bar{y})$ 的数值会更大，协方差也随之变大，而事实上二者的相关关系并无变化。为了避免这种情况，我们将使用**相关系数**（correlation coefficient）对两变量间的相关关系进行量度。

3.5.3 相关系数

对于样本数据，皮尔逊积矩相关系数的定义如下。

皮尔逊积矩相关系数：样本数据

$$r_{xy}=\frac{s_{xy}}{s_x s_y} \tag{3-14}$$

式中，r_{xy} 为样本相关系数；s_{xy} 为样本协方差；s_x 为 x 的样本标准差；s_y 为 y 的样本标准差。

⊖ 协方差是两变量线性关系的度量。

式（3-14）表明样本数据的皮尔逊积矩相关系数（通常简称为样本相关系数）的计算方法是：用样本协方差除以 x 的标准差与 y 的标准差的乘积。

现在让我们来计算音像设备商店的样本相关系数。使用表3-6中的数据，我们可以得到两个变量的样本标准差。

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{20}{9}} = 1.49$$

$$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{566}{9}} = 7.93$$

现在，由于 $s_{xy} = 11$，我们可以得到样本相关系数为

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{1.49 \times 7.93} = +0.93$$

总体相关系数用希腊字母 ρ_{xy} 表示，其计算公式如下。

皮尔逊积矩相关系数：总体数据

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3-15}$$

式中，ρ_{xy} 为总体相关系数；σ_{xy} 为总体协方差；σ_x 为 x 的总体标准差；σ_y 为 y 的总体标准差。

样本相关系数[⊖] r_{xy} 给出了总体相关系数 ρ_{xy} 的一个估计。

3.5.4　样本相关系数的解释

首先让我们考虑一个简单的样本，以说明完全正线性相关的概念。图3-12的散点图描述了下列样本数据 x 和 y 之间的关系。

x_i	5	10	15
y_i	10	30	50

经过这三个点的直线显示，在 x 和 y 之间存在完全线性关系。为了应用式（3-14）来计算样本相关系数，我们必须先计算 s_{xy}，s_x 和 s_y。在表3-8中列出了一些计算过程。利用表3-8中的结果，我们得到

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{200}{2} = 100$$

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{50}{2}} = 5$$

$$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{800}{2}} = 20$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5 \times 20} = 1$$

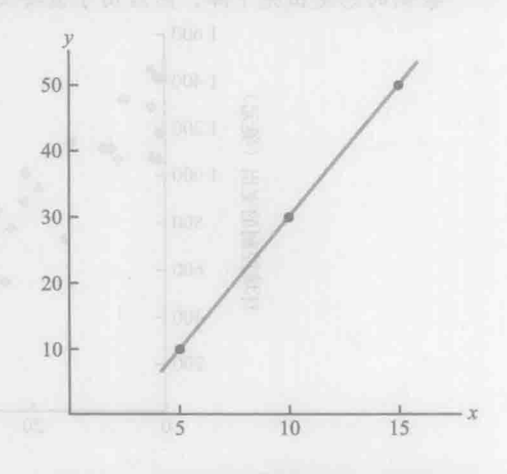


图3-12　描述完全正线性关系的散点图

于是，我们看到样本相关系数的值为1。

⊖ 样本相关系数 r_{xy} 是总体相关系数 ρ_{xy} 的点估计量。

表 3-8 样本相关系数的计算

	x_i	y_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
	5	10	-5	25	-20	400	100
	10	30	0	0	0	0	0
	15	50	5	25	20	400	100
合计	30	90	0	50	0	800	200
	$\bar{x}=10$	$\bar{y}=30$					

一般地，我们发现如果数据集中所有的点都在一条斜率为正的直线上，则样本相关系数[⊖]的值为 +1；也就是说，样本相关系数的值为 +1 对应于 x 和 y 之间存在一个完全正线性关系。如果数据集中所有的点都在一条斜率为负的直线上，则样本相关系数的值为 -1；也就是说，样本相关系数的值为 -1 对应于 x 和 y 之间存在一个完全负线性关系。

现在我们假设，某个数据集表示 x 和 y 存在一个正的线性关系，但不是完全线性的。r_{xy}的值将小于1，表示在散点图上点不全在一条直线上。当随着数据集中的点越来越偏离完全正线性关系时，r_{xy}的值就变得越来越小。当 r_{xy}等于零时，表明 x 和 y 之间不存在线性关系，r_{xy}的值接近于零则表明弱的线性关系。

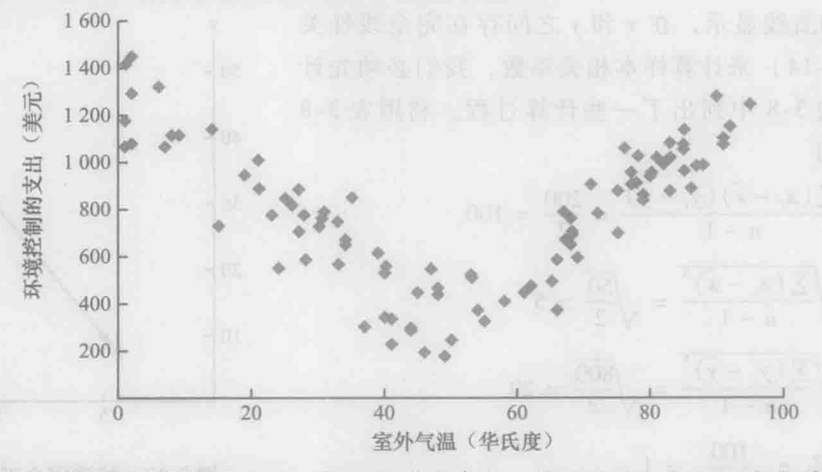
对于音像设备商店的数据，我们知道 $r_{xy}=0.93$。因此，我们可以得出广告次数和销售额之间存在着强的线性关系。更具体地说，广告次数的增加会带来销售额的增加。

最后，我们注意相关系数提供了线性但不是因果关系的一个度量。两个变量之间较高的相关系数，并不意味着一个变量的变化会引起另一个变量的变化。例如，我们看到，饭店的质量等级和代表性餐价是正相关的。但是，简单地增加饭店的餐价并不会提供单调的质量等级。

注释和评论

1. 由于相关系数仅仅度量两个数量变量之间线性相关关系的强度，因此，当两个变量之间的关系是非线性时，相关系数可能接近于0，表明没有线性关系。例如，下面的散点图表明在过去100天中室外的日最高气温和受环境控制（供暖和制冷）的一个小零售商店总支出的关系。

   这些数据的样本相关系数 $r_{xy}=-0.007$，表明两个变量之间没有线性相关关系。但是，散点图给出了非线性关系的很直观证据。因此，我们可以看到随着室外日最高气温的增加，由于需要少量暖气，受环境控制的总支出先下降，然后由于需要大量制冷，总支出增加。



2. 虽然在评价两个数量变量的相关关系时，相关系数很有用，其他度量——如斯皮尔曼等级相关系数——可用于评价至少有一个变量是名义或顺序变量的两个变量的相关系数。我们将在第18章讨论斯皮尔曼等级相关系数。

⊖ 相关系数的范围是 -1 ~ +1。当相关系数接近于 -1 或 +1 时，表示强的线性关系，而相关系数越接近于0，线性关系也越弱。

练习

方法

56. 下面是两个变量的5次观测值

x_i	6	11	15	21	27
y_i	6	9	6	17	12

a. 绘制这些数据的散点图。

b. 散点图中表明x和y之间存在何种关系？

c. 计算并解释样本协方差。

d. 计算并解释样本相关系数。

应用

58. 运输部门对中型汽车的行驶速度和油耗进行了研究，得到下面的数据。

行驶速度（英里/小时）	30	50	40	55	30	25	60	25	50	55
油耗（英里/加仑）	28	25	25	23	30	32	21	35	26	25

计算并解释样本相关系数。

60. 道琼斯工业平均指数（DJIA）和标准普尔500指数（S&P）是用来度量股票市场的业绩的两个指数。DJIA是以30家大公司的股票价格为依据；S&P 500是以500家公司的股票价格为依据。如果用DJIA和S&P 500度量股票市场的业绩，它们相关吗？由为期三个月的9天组成一个样本，下面是DJIA和S&P500每天上涨和下跌的百分比样本数据（《华尔街日报》，2006年1月15日~3月10日）。

DJIA	0.20	0.82	−0.99	0.04	−0.24	1.01	0.30	0.55	−0.25
S&P 500	0.24	0.19	−0.91	0.08	−0.33	0.87	0.36	0.83	−0.16

a. 绘制散点图。

b. 计算这些数据的样本相关系数。

c. 讨论DJIA和S&P500之间的关系。在得到每天股票市场业绩的一般印象之前，你是否需要审核这两个指数？

3.6　数据仪表板：增加数值度量以提高有效性

在2.5节，我们给出了用于描述汇总和表述数据集信息的图形显示效用的数据可视化术语的简介。数据可视化的目的是尽可能有效和清晰地传递数据的重要信息。数据可视化工具使用最广泛的一种是数据仪表板。数据仪表板是一个直观显示的集合，这个直观显示以易于阅读、了解和解释的方式组织和表述用于监控公司或机构业绩的信息。本节，我们演示数值度量的增加如何能提高显示整体的有效性，来延伸数据仪表板的讨论。

关键表现指标（KPIs）（如平均数和标准差）数值度量的增加，对数据仪表板来说至关重要，因为数值度量常常提供KPIs评估的基准和目标。另外，数据仪表板也常常包括含有数值度量的图形显示，将数值度量作为显示的一个组成部分。我们必须牢记，数据仪表板的用途是以易于阅读、了解和解释的方式提供KPIs信息。增加数值度量和利用数值度量的图表可以帮助我们实现这些目标。

为了说明数据仪表板中数值度量的使用，回顾为了介绍数据仪表板的概念，在2.5节中我们使用的Grogan石油公司的应用。Grogan石油公司在得克萨斯州的奥斯汀（其总部所在）、休斯敦和达拉斯这三个城市设有办事处。Grogan石油公司位于奥斯汀办事处的信息技术（IT）呼叫中心处理员工关于软件、互联网和电子邮件问题等与计算机相关问题的呼叫。绘制监控呼叫中心表现的数据仪表板如图3-13所示。这个仪表板的重要组成部分如下：

- 仪表板左上角的结构条形图展示过去时间内每一类问题（软件、互联网和电子邮件）的呼叫次数。
- 仪表板右上角的饼形图显示呼叫中心员工在每一类问题或没有呼叫工作（空闲）所用时间的比例。
- 对每一个超过15分钟之前收到的未被解决的案例，仪表板中间左侧的条形图展示了这些未被解决的案例中每一个的时间长度。
- 仪表板中间右侧的条形图展示了办事处（休斯敦、达拉斯、奥斯汀）的每一类问题的呼叫次数。
- 仪表板底部的直方图展示了当前班次所有被解决的案例所用时间的分布。

为了获得呼叫中心业绩的更多内涵，Grogan石油公司的信息技术经理决定扩充当前的仪表板，增加收到每一类问题（电子邮件、互联网和软件）呼叫需要解决时间的箱形图。另外，在仪表板左下部分增加展示解决单个案例所用时间的图形。最后，信息技术经理增加每一类问题的汇总统计量和本班次前几小时中每小时的汇总统计量的显示。更新后的仪表板如图3-14所示。

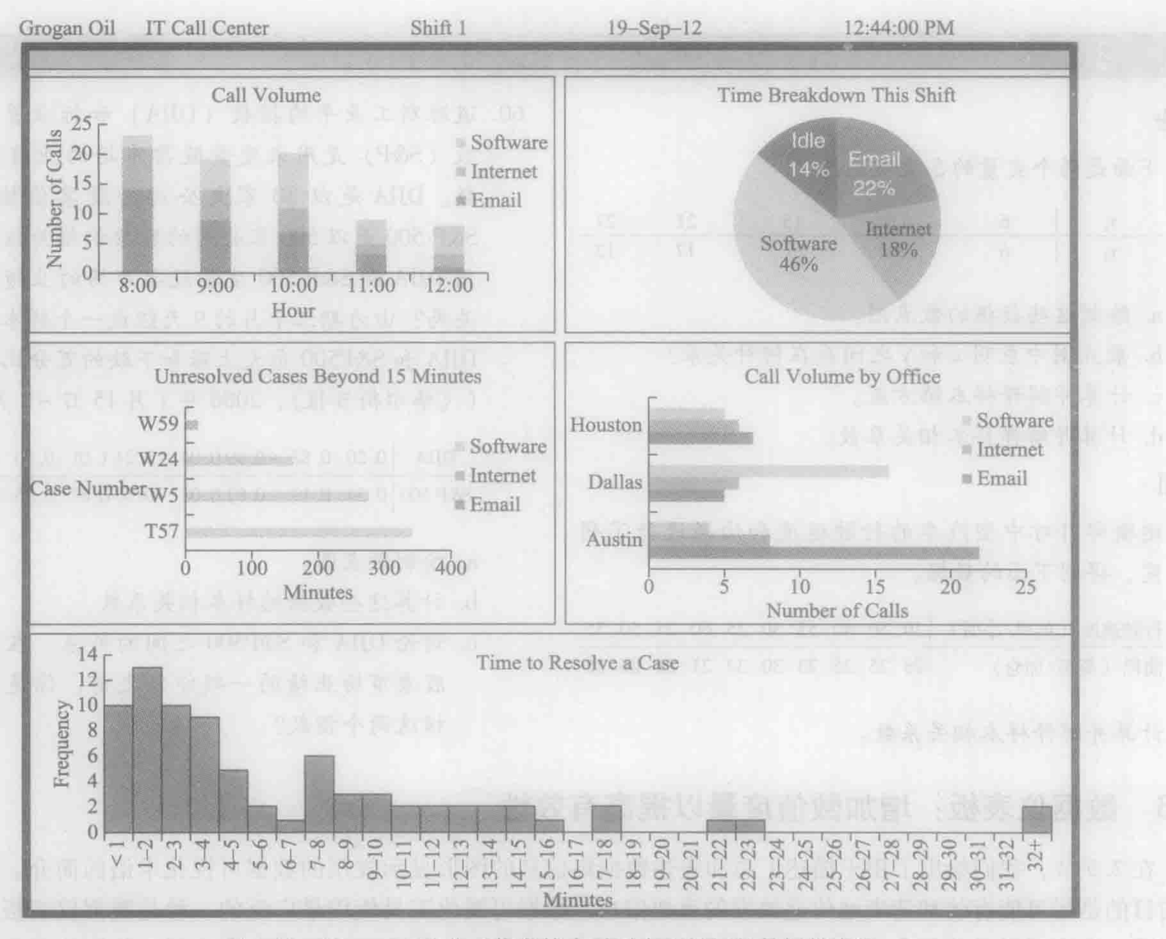


图 3-13 Grogan 石油公司信息技术呼叫中心最初的数据仪表板

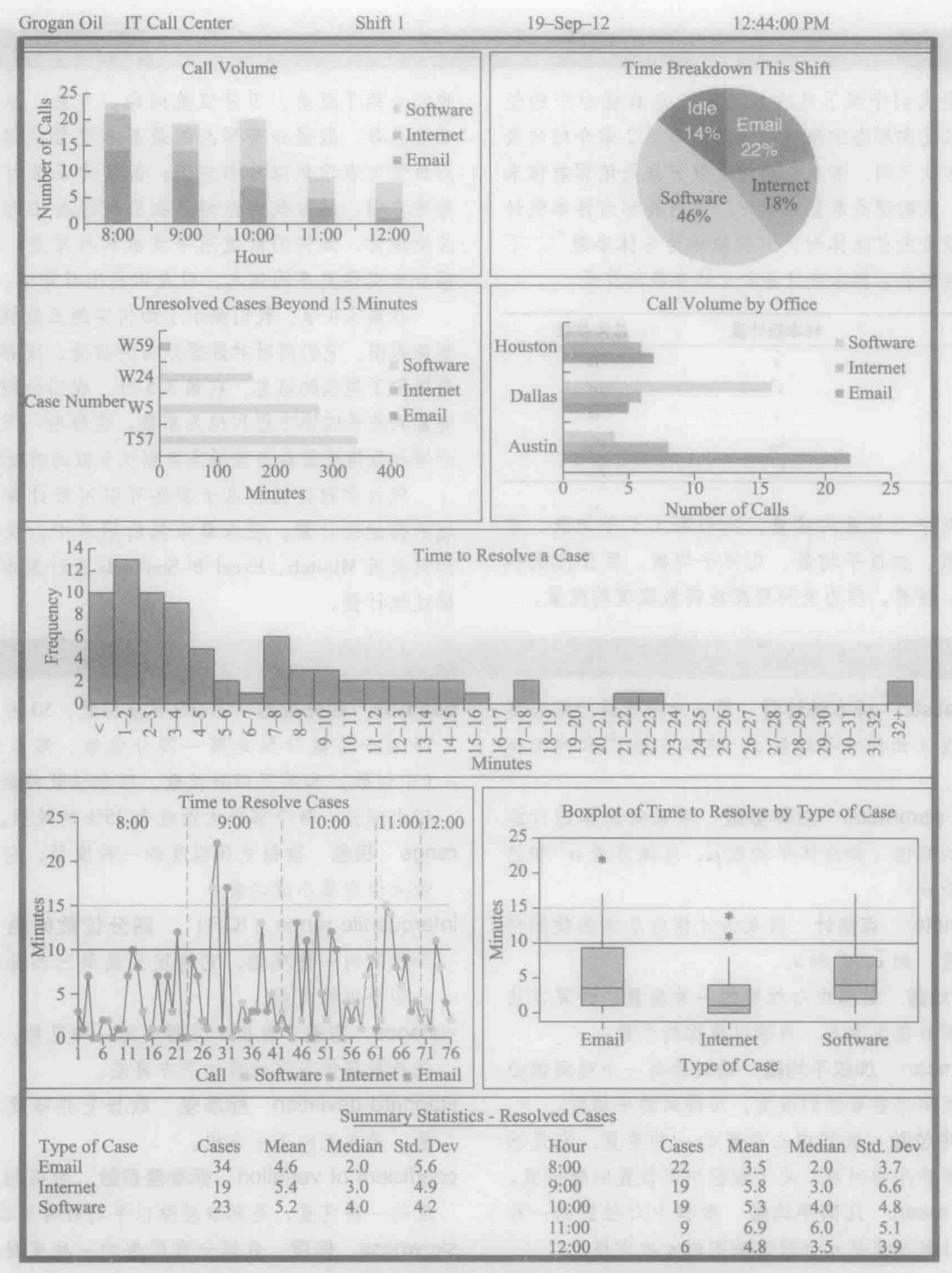
信息技术中心经理制定平均 10 分钟解决一个案例的目标业绩水平或基准。此外，中心决定一个案例的解决时间超过 15 分钟是不合格的。为了反映这些基准，在解决单个案例所用时间的图和收到每一类问题呼叫需要解决时间箱形图这两张图上，都增添了平均目标值 10 分钟的黑色横线和最大可接受水平 15 分钟的红色横线。

图 3-14 仪表板中的汇总统计量展示：一个电子邮件案例的平均解决时间为 4.6 分钟，一个互联网案例的平均解决时间为 5.4 分钟，以及一个软件案例的平均解决时间为 5.2 分钟。因此，每一类案例的平均解决时间都比目标平均数（10 分钟）要好。

查看箱形图，我们看到电子邮件案例对应的箱形图比其他两类案例对应的箱形图要“大”。汇总统计量也显示电子邮件案例解决时间的标准差比其他类型案例解决时间的标准差大。这让我们在两张新图中仔细查看电子邮件案例。电子邮件案例的箱形图有一条延伸到 15 分钟之外的触须线，且有一个大于 15 分钟的异常值。解决单个案例所用时间图（在仪表板左下部分）展示在 9:00 时段由于有两个电子邮件呼叫解决时间比目标最大时间（15 分钟）要长。

这个分析使信息技术呼叫中心经理进一步审查为什么电子邮件案例解决时间比互联网或软件案例有更大变异性。基于这些分析，信息技术经理还可能决定审查导致这两个超过 15 分钟才被解决的电子邮件案例异常大的解决时间的工作情况。

解决单个案例所用时间图还显示当前班次第一小时时段内收到的许多呼叫解决相对较快；图形还显示整个上午案例的解决时间逐渐增加。这可能是由于在本班次稍后时间有生产复杂问题的趋势，或随时间推移而堆积的呼叫积压。尽管汇总统计量暗示在 9:00 时段提交的案例解决时间较长，解决单个案例所用时间图显示在这个时段上报告有两个费时的电子邮件案例和一个费时的软件案例，这可以解释 9:00 时段为什么比本班次其他时段案例的平均解决时间要长。总之，本班次报告的案例一般在 15 分钟或者更少的时间内被解决。



Type of Case	Cases	Mean	Median	Std. Dev
Email	34	4.6	2.0	5.6
Internet	19	5.4	3.0	4.9
Software	23	5.2	4.0	4.2

Hour	Cases	Mean	Median	Std. Dev
8:00	22	3.5	2.0	3.7
9:00	19	5.8	3.0	6.6
10:00	19	5.3	4.0	4.8
11:00	9	6.9	6.0	5.1
12:00	6	4.8	3.5	3.9

图 3-14　Grogan 石油公司信息技术呼叫中心更新的数据仪表板

如 Grogan 石油公司的数据仪表板这类仪表板常常是互动的[⊖]。例如，当经理用鼠标或触摸屏在显示上将光标定位或指向显示上的某事，如解决问题的时间、收到呼叫的时间，以及个人和所报告问题的位置这些附加信息将会出现。点击单个项目也可以在个别情况下，带给用户一个新的分析水平。

⊖ 数据仪表板的互动功能定义为深化，它允许用户存取信息，并且越来越详细分析。

小结

本章里我们介绍了几种能用来汇总数据分布的位置、变异程度和形态的描述统计量。与第2章介绍的表格和图形方法不同，本章介绍的量度方法是依据数值来汇总数据。当数值是来自样本时，它们被称为样本统计量；当数值是来自总体时，它们被称为总体参数[⊖]。下面是一些用来表示样本统计量和总体参数的符号。

	样本统计量	总体参数
平均数	$\bar{x}$	μ
方差	s^2	σ^2
标准差	s	σ
协方差	s_{xy}	σ_{xy}
相关系数	r_{xy}	ρ_{xy}

作为对中心位置的度量，我们定义了平均数、中位数、众数、加权平均数、几何平均数、百分位数和四分位数。接着，作为变异程度或离散程度的度量，我们介绍了极差、四分位数间距、方差、标准差和标准差系数。数据分布形态的最基本度量是偏度。偏度为负值时表示数据分布左偏；偏度为正值时表示数据分布右偏。然后我们介绍了在应用经验法则和切比雪夫定理时，如何同时使用平均数和标准差，从而对数据分布提供更多的信息，以及识别出异常值。

在第3.4节，我们演示了如何实施五数概括法和绘制箱形图，它们同时对数据分布的位置、变异程度和形态提供了类似的信息。在第3.5节，我们介绍了度量两变量间关系的协方差和相关系数。在最后一节，我们演示增加数值度量如何能提高数据仪表板的有效性。

统计学软件包和电子表格可以用来计算我们讨论过的描述统计量。在本章末尾的附录中，我们演示了如何使用Minitab、Excel和StatTools来计算本章介绍的描述统计量。

关键术语

sample statistic 样本统计量 用来对样本进行综合度量的数值（如样本平均数$\bar{x}$，样本方差s^2和样本标准差s）。

population parameter 总体参数 用来对总体进行综合度量的数值（如总体平均数μ，总体方差σ^2和总体标准差σ）。

point estimator 点估计 用来估计相应总体参数的样本统计量，如$\bar{x}$，s^2和s。

mean 平均数 数据中心位置的一种度量，计算方法是将所有数据值加总，再除以数据的个数。

weighted mean 加权平均数 通过给每一个观测值分配一个反映其重要性的权重，而得到的平均数。

median 中位数 数据中心位置的一种度量，它是当数据按照升序排列时，处于数据中间位置的数据值。

geometric mean 几何平均数 数据中心位置的一种度量，计算方法是n个数值乘积的n次方根。

mode 众数 数据的位置的一种度量，被定义为是出现次数最多的数值。

percentile 百分位数 一个数值，在数据集中至少有$p\%$的观测值小于或等于该值，且至少有$(100-p)\%$的观测值大于或等于该值。第50百分位数就是中位数。

quartiles 四分位数 第25百分位数，50百分位数和75百分位数分别是第一四分位数、第二四分位数（中位数）和第三四分位数。四分位数将数据集分为四个部分，每个部分大约包含25%的数据。

range 极差 数据变异程度的一种度量，它的定义是最大值与最小值之差。

interquartile range (IQR) 四分位数间距 数据变异程度的一种度量，它的定义是第三四分位数与第一四分位数之差。

variance 方差 数据变异程度的一种度量，计算的依据是数据值与平均数的平方离差。

standard deviation 标准差 数据变异程度的一种度量，是方差的正平方根。

coefficient of variation 标准差系数 数据相对变异程度的一种度量，是标准差除以平均数再乘以100。

skewness 偏度 数据分布形态的一种度量。数据左偏的偏度为负值；数据分布对称的偏度为0；数据右偏的偏度为正值。

***z*-score *z*-分数** x_i与平均数的离差$(x_i-\bar{x})$除以标准差s得到的数值。z-分数也称为标准化数值，表示以标准差的个数度量的x_i与平均数的距离。

⊖ 在统计推断中，样本统计量被称为总体参数的点估计。

Chebyshev's theorem　切比雪夫定理　用来陈述与平均数的距离在一个指定个数的标准差之内的数据值所占比例的定理。

empirical rule　经验法则　对于具有钟形分布的数据，可用来计算与平均数的距离在1个、2个和3个标准差之内的数据值所占比例的法则。

outlier　异常值　小得不正常或大得不正常的数值。

five-number summary　五数概括法　一种用五个数值：最小值、第一四分位数、中位数、第三四分位数和最大值来汇总数据技术。

box plot　箱形图　以五数概括法为依据的一种数据的图形汇总方法。

covariance　协方差　两变量间线性关系的一种度量。正值表示正的关系，负值表示负的关系。

correlation of coefficient　相关系数　两变量间线性关系的一种度量，其取值在 -1 到 +1 之间。接近 +1 的值表示强的正线性关系；接近 -1 的值表示强的负线性关系；接近于零的值表示没有线性关系。

重要公式

样本平均数

$$\bar{x} = \frac{\sum x_i}{n} \tag{3-1}$$

总体平均数

$$\mu = \frac{\sum x_i}{N} \tag{3-2}$$

加权平均数

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \tag{3-3}$$

几何平均数

$$\bar{x}_g = \sqrt[n]{x_1 x_2 \cdots x_n} = [x_1 x_2 \cdots x_n]^{1/n} \tag{3-4}$$

四分位数间距

$$IQR = Q_3 - Q_1 \tag{3-5}$$

总体方差

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \tag{3-6}$$

样本方差

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{3-7}$$

标准差

$$\text{样本标准差 } s = \sqrt{s^2} \tag{3-8}$$

$$\text{总体标准差 } \sigma = \sqrt{\sigma^2} \tag{3-9}$$

标准差系数

$$\left(\frac{\text{标准差}}{\text{平均数}} \times 100\right)\% \tag{3-10}$$

z-分数

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3-11}$$

样本协方差

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{3-12}$$

总体协方差

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N} \tag{3-13}$$

皮尔逊积矩相关系数：样本数据

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3-14}$$

皮尔逊积矩相关系数：总体数据

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3-15}$$

补充练习

62. 美国人一周中外出就餐的平均次数从2008年的4.0次下降到2012年的3.8次（Zagat.com，2012年4月1日）。一个由20个家庭组成的样本，上周他们外出就餐次数的数据如下。

6	1	5	3	7	3	0	3	1	3
4	1	2	4	1	0	5	6	3	1

a. 计算平均数和中位数。
b. 计算第一四分位数和第三四分位数。
c. 计算极差和四分位数间距。
d. 计算方差和标准差。
e. 这些数据的偏度为0.34。对这个分布的形态进行评论。是你期望的形态吗？为什么？
f. 数据中包含异常值吗？

64. 病人在艾尔帕索医生诊所的平均等候时间刚刚超过29分钟，而美国全国的平均候诊时间为21分钟。事实上，艾尔帕索有全美国最长的候诊时间（EL Paso Time，2012年1月8日）。为了处理病人长时间等候问题，一些医生诊所使用了等待追踪系

统来告知病人预计等候时间。根据这些信息，病人可以调整他们的到达时间，在等候室花费少量的时间。下面是一个没有等待追踪系统的诊所样本和一个有等待追踪系统的诊所样本，它们病人的候诊时间数据（单位：分钟）。

没有等待追踪系统	24	67	17	20	31	44	12	23	16	37
有等待追踪系统	31	11	14	18	12	37	9	13	12	15

a. 对于有等待追踪系统的诊所，病人候诊时间的平均数和中位数是多少？对于没有等待追踪系统的诊所，病人候诊时间的平均数和中位数是多少？

b. 对于有等待追踪系统的诊所，病人候诊时间的方差和标准差是多少？对于没有等待追踪系统的诊所，病人候诊时间的方差和标准差是多少？

c. 有等待追踪系统的诊所病人的候诊时间比没有等待追踪系统的诊所短吗？请解释。

d. 只考虑没有等待追踪系统的诊所，样本中第10位病人的z-分数是多少？

e. 只考虑有等待追踪系统的诊所，样本中第6位病人的z-分数是多少？这个z-分数如何与（d）中计算z-分数相比较？

f. 基于z-分数，没有等待追踪系统的诊所数据包含异常值吗？基于z-分数，有等待追踪系统的诊所数据包含异常值吗？

66. 小型企业主往往期望薪酬服务公司来处理其员工的工资问题。原因是小型企业主主要面临复杂的税收条例以及对就业交税错误的昂贵罚款。根据美国国税局的报告，有26%的小型企业主的就业纳税申报单有错误，从而导致对企业主的税务罚款（《华尔街日报》，2006年1月30日）。20个小型企业主税务罚款的样本数据如下。

820	270	450	1 010	890	700	1 350	350	300	1 200
390	730	2 040	230	640	350	420	270	370	620

a. 对于不正确的就业纳税申报单，平均税务罚款是多少？

b. 标准差是多少？

c. 最高的罚款2 040美元是异常值吗？

d. 小型企业主雇用薪酬服务公司来处理包括其就业纳税申报单等员工的工资问题，有哪些优点？

68. 美国房地产协会报道了美国房屋价格的中位数和5年期间房屋价格中位数的增长率（《华尔街日报》，2006年1月16日）。利用下面房屋价格（单位：1 000美元）的样本数据回答下列问题。

995.9	48.8	175.0	263.5	298.0	218.9	209.0
628.3	111.0	212.9	92.6	2 325.0	958.0	212.5

a. 房屋价格样本中位数是多少？

b. 在2001年1月，美国房地产协会报道了美国房屋价格的中位数为139 300美元。5年期间房屋价格中位数增长率是多少？

c. 样本数据的第一四分位数和第三四分位数是多少？

d. 对房屋价格应用五数概括法。

e. 数据中有异常值吗？

f. 房屋价格的样本均值是多少？为什么美国房地产协会在报道中更喜欢使用房屋价格的中位数？

70. *Travel + Leisure* 杂志提供世界上500家最佳酒店的年度一览表（*Travel + Leisure*，2009年1月）。杂志给出了每一家酒店的排名，同时还对每一家酒店的规模、环境、双人间的每晚价格进行了简单的描述。美国12家顶级酒店的样本如下。

酒店	地点	客房数	价格/晚
Boulders Resort & Spa	菲尼克斯，亚利桑那州	220	499
Disney's Wilderness Lodge	奥兰多，佛罗里达州	727	340
Four Seasons Hotel Beverly Hills	洛杉矶，加利福尼亚州	285	585
Four Seasons Hotel	波士顿，马萨诸塞州	273	495
Hay-Adams	华盛顿，特区	145	495
Inn on Biltmore Estate	阿什维尔，北卡罗来纳州	213	279
Loews Ventana Canyon Resort	菲尼克斯，亚利桑那州	398	279
Mauna Lani Bay Hotel	夏威夷岛	343	455
Montage Laguna Beach	Laguna海滩，加利福尼亚州	250	595
Sofitel Water Tower	芝加哥，伊利诺伊州	414	367
St. Regis Monarch Beach	达纳半岛，加利福尼亚州	400	675
The Broadmoor	科罗拉多斯普林斯，科罗拉多州	700	420

a. 客房的平均数是多少？

b. 双人间每晚价格的平均数是多少？

c. 以客房数为横轴，每晚价格为纵轴绘制散点图。客房数和每晚价格之间相关吗？请讨论。

d. 样本相关系数是多少？关于客房数和双人间每晚价格之间的关系，这个相关系数能告诉你怎样的信息？这个结论合理吗？请讨论。

72. 职业棒球大联盟的球队春训期间的记录能说明球队在常规赛中的表现吗？在过去6年期间，球队在春训期间的获胜率与它在常规赛中的获胜率之间的相关系数为0.18。(《华尔街日报》，2009年3月30日)。下面是2008赛季14支美国职业棒球联盟球队获胜率数据。

球队	春训	常规赛	球队	春训	常规赛
巴尔的摩 Orioles	0.407	0.422	明尼苏达 Twins	0.500	0.540
波士顿 Red Sox	0.429	0.586	纽约 Yankees	0.577	0.549
芝加哥 White Sox	0.417	0.546	奥克兰 A's	0.692	0.466
克利夫兰 Indians	0.569	0.500	西雅图 Mariners	0.500	0.377
底特律 Tigers	0.569	0.457	坦帕 Bay Rays	0.731	0.599
堪萨斯 City Royals	0.533	0.463	得克萨斯 Rangers	0.643	0.488
洛杉矶 Angels	0.724	0.617	多伦多 Blue Jays	0.448	0.531

a. 春训和常规赛获胜率之间的相关系数是多少？

b. 对一支球队春训期间的记录说明球队在常规赛中的表现，你有什么结论？会出现这种情况的原因是什么？请讨论。

74. 在一个限速为每小时55英里的公路上行驶的汽车会受到州警察雷达系统的车速检测。下面一个是车速的频数分布。

车速（英里/小时）	频数
45～49	10
50～54	40
55～59	150
60～64	175
65～69	75
70～74	15
75～79	10
合计	475

a. 在这条公路上行驶的汽车的平均车速是多少？

b. 计算方差和标准差。

案例3-1　Pelican 商店

Pelican 商店是 National Clothing 的一个分支，是一家在全美范围内经营妇女服饰的连锁商店。最近，它举办了一项促销活动，向其他 National Clothing 商店的顾客赠送优惠券。在促销活动期内的某一天，Pelican 商店抽取了100名持信用卡交易的顾客组成一个样本，搜集到的数据存在名为 PelicanStores 的文件中。表3-9是数据集的一部分。Proprietary Card 付款方法是指使用 National Clothing 签账卡收费。使用优惠券购物的顾客定义为促销顾客，没有使用优惠券购物的顾客定义为普通顾客。因为优惠券不会派发给 Pelican 商店的普通顾客，管理者认为，持促销优惠券的顾客生产的销售额与其他顾客不同。当然，Pelican 的管理者也希望促销顾客会继续在他的商店购物。

表3-9中的大多数变量不需要解释，但有两个变量需要稍做说明。

项目　购买商品的总件数

净销售额　信用卡支付的总金额（单位：美元）

Pelican 的管理者希望使用这些样本数据，以了解其顾客的基本情况并对使用优惠券的促销活动进行评估。

表3-9　Pelican 商店的100名持信用卡消费的样本数据

顾客	顾客类型	项目	净销售额	支付方法	性别	婚姻状况	年龄
1	常规性	1	39.50	Discover	男	已婚	32
2	奖励性	1	102.40	Proprietary Card	女	已婚	36
3	常规性	1	22.50	Proprietary Card	女	已婚	32
4	奖励性	5	100.40	Proprietary Card	女	已婚	28
5	常规性	2	54.00	MasterCard	女	已婚	34
6	常规性	1	44.50	MasterCard	女	已婚	44
7	奖励性	2	78.00	Proprietary Card	女	已婚	30
8	常规性	1	22.50	Visa	女	已婚	40

（续）

顾客	顾客类型	项目	净销售额	支付方法	性别	婚姻状况	年龄
9	奖励性	2	56.52	Proprietary Card	女	已婚	46
10	常规性	1	44.50	Proprietary Card	女	已婚	36
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
96	常规性	1	39.50	MasterCard	女	已婚	44
97	奖励性	9	253.00	Proprietary Card	女	已婚	30
98	奖励性	10	287.59	Proprietary Card	女	已婚	52
99	奖励性	2	47.60	Proprietary Card	女	已婚	30
100	奖励性	1	28.44	Proprietary Card	女	已婚	44

管理报告

使用本章介绍的描述统计的方法来汇总数据并评价你的发现。你的报告至少应该包括以下内容：

1. 净销售额的描述统计量和顾客的各种不同分类的净销售额的描述统计量。
2. 关于年龄与净销售额之间关系的描述统计量。

案例3-2 电影业

电影业是一个竞争激烈的行业，每年有超过50个制片厂制作出几百部新电影，每部电影商业上的成功差异很大。周末首映票房收入（单位：100万美元）、票房总收入（单位：100万美元）、放映电影的剧院数，以及电影放映的周数是衡量一部电影是否成功的最常用的变量。2011年制作的100部大电影的样本数据存在文件名为Movies文件中（Box Office Mojo，2012年3月17日）。表3-10是文件中前10部的电影的有关数据。注意，有些电影，如《战马》在2011年年底上映，2012年继续放映。

表3-10 10部电影的业绩数据

电影	首映票房（100万美元）	票房总收入（100万美元）	剧院数	放映周数
哈利波特与死亡圣器第二部	169.19	381.01	4 375	19
变形金刚：月黑之时	97.85	352.39	4 088	15
暮光之城：破晓第一部	138.12	281.29	4 066	14
宿醉2	85.95	254.46	3 675	16
加勒比海盗：惊涛怪浪	90.15	241.07	4 164	19
速度与激情5	86.20	209.84	3 793	15
碟中谍4：幽灵协议	12.79	208.55	3 555	13
赛车总动员2	66.14	191.45	4 115	25
福尔摩斯：诡影游戏	39.64	186.59	3 703	13
雷神	65.72	181.03	3 963	16

管理报告

使用本章介绍的描述统计的方法来了解这些变量对一部电影成功有怎样的贡献。你的报告应该包括以下内容：

1. 四个变量中每个变量的描述统计量，接着对每个描述统计量得出的关于电影业的情况进行讨论。
2. 如果有，哪些电影被认为是优异表现的异常值？请解释。
3. 列出总票房收入与每一个其他变量之间关系的描述统计量，请解释。

案例3-3 亚太地区的商学院

当前追求工商管理专业较高的学历已是一种国际性的潮流。有调查表明，越来越多的亚洲人选择攻读工商管理硕士（MBA）学位，把它作为通向企业成功的一种途径。因此，亚太地区商学院MBA课程的申请者人数持续增加。

在整个亚太地区，有成千上万的亚洲人暂时搁置自己的事业，并花两年的时间来追求系统的工商管理教育。这些工商管理课程众所周知是十分繁重的，包括经济学、金融学、市场营销、行为科学、劳动关系、决策论、战略思想、经济法等。表3-11的数据集中列出了亚太地区一流商学院的一些情况。

管理报告

利用描述统计学的方法来汇总表3-11中的数据，并讨论你的发现。

1. 对数据集中的每一个变量进行汇总。根据最大值、最小值、适当的平均数和分位数，进行评价和解释。对于亚太地区的商学院，这些描述统计量能够提供什么新的见解？
2. 汇总数据以进行下列比较：
   a. 本国学生和国外学生学费的差别。
   b. 要求工作经验和不要求工作经验的学校学生平均起薪的差别。
   c. 要求英语测试和不要求英语测试的学校学生的起薪的差别。
3. 起薪与学费有关吗？
4. 对表3-11的数据作出其他的图形和数值汇总，以便同他人交流。

表3-11 亚太地区25所商学院的数据

商学院名称	录取名额	每个学院人数	本国学生学费（美元）	外国学生学费（美元）	年龄	国外学生比例（%）	是否要求GMAT	是否要求英语测试	是否要求工作经验	起薪（美元）
墨尔本商学院	200	5	24 420	29 600	28	47	是	否	是	71 400
新南威尔士大学（悉尼）	228	4	19 993	32 582	29	28	是	否	是	65 200
印度管理学院（阿默达巴德）	392	5	4 300	4 300	22	0	否	否	否	7 100
香港中文大学	90	5	11 140	11 140	29	10	是	否	否	31 000
日本国际大学（新潟）	126	4	33 060	33 060	28	60	是	是	否	87 000
亚洲管理学院（马尼拉）	389	5	7 562	9 000	25	50	是	否	是	22 800
印度管理学院（班加罗尔）	380	5	3 935	16 000	23	1	是	否	否	7 500
新加坡国立大学	147	6	6 146	7 170	29	51	是	是	是	43 300
印度管理学院（加尔各答）	463	8	2 880	16 000	23	0	否	否	否	7 400
澳大利亚国立大学（堪培拉）	42	2	20 300	20 300	30	80	是	是	是	46 600
南洋理工大学（新加坡）	50	5	8 500	8 500	32	20	是	否	是	49 300
昆士兰大学（布里斯班）	138	17	16 000	22 800	32	26	否	否	是	49 600
香港科技大学	60	2	11 513	11 513	26	37	是	否	是	34 000
麦夸里管理研究生院（悉尼）	12	8	17 172	19 778	34	27	否	否	是	60 100
Chulalongkorn大学（曼谷）	200	7	17 355	17 355	25	6	是	否	是	17 600
Monash Mt. Eliza商学院（墨尔本）	350	13	16 200	22 500	30	30	是	是	是	52 500
亚洲管理学院（曼谷）	300	10	18 200	18 200	29	90	否	是	是	25 000
阿德莱德大学	20	19	16 426	23 100	30	10	否	否	是	66 000
梅西大学（新西兰，北帕默斯顿）	30	15	13 106	21 625	37	35	否	是	是	41 400
皇家墨尔本技术工商学院	30	7	13 880	17 765	32	30	否	是	是	48 900
Jamnalal Bajaj管理研究院（孟买）	240	9	1 000	1 000	24	0	否	否	是	7 000
柯廷理工学院（珀思）	98	15	9 475	19 097	29	43	是	否	是	55 000
拉合尔管理科学学院	70	14	11 250	26 300	23	2.5	否	否	否	7 500
马来西亚Sains大学（槟城）	30	5	2 260	2 260	32	15	否	是	是	16 000
De La Salle大学（马尼拉）	44	17	3 300	3 600	28	3.5	是	否	是	13 100

案例 3-4 天使巧克力的网络交易

天使巧克力在位于纽约州萨拉托加的工厂和零售店制造和销售优质巧克力产品。两年前公司建立了网站，并开始在互联网上销售其产品。网络销售超过了公司的期待，管理者现在考虑增加未来销售额的战略。为了了解网络消费者的更多信息，从前几个月的销售中，选择了50次天使巧克力交易组成一个样本，数据中包括每笔交易的发生日期、消费者使用的浏览器类型、在网站上度过的时间、观看网页的数量以及每名消费者的消费金额，数据存在名为 Shoppers 的文件中。表3-12是数据的一部分。

表 3-12 一个50笔天使巧克力网络交易的样本数据

消费者	日期	浏览器	时间（分钟）	观看网页	消费金额（美元）
1	周一	因特网	12.0	4	54.52
2	周三	其他	19.5	6	94.90
3	周一	因特网	8.5	4	26.68
4	周二	火狐	11.4	2	44.73
5	周三	因特网	11.3	4	66.27
6	周六	火狐	10.5	6	67.80
7	周日	因特网	11.4	2	36.04
⋮	⋮	⋮	⋮	⋮	⋮
48	周五	因特网	9.7	5	103.15
49	周一	其他	7.3	6	52.15
50	周五	因特网	13.4	3	98.75

天使巧克力希望用这些样本数据，来确定在线购物者是否花费更多时间、观看更多网页以及他们在浏览网站时是否消费更多。公司也希望调查销售日期和浏览器类型的效果。

管理报告

使用描述统计学的方法来了解访问天使巧克力网站的消费者信息。你的报告应该包括下面的内容：

1. 购物者在网站上度过的时间长度、观看网页的数量，以及每笔交易的平均消费金额的图表和数值汇总。从这些数值汇总中，讨论你了解的关于天使巧克力在线购物者的情况。

2. 对每周的交易日期汇总频数、消费总额以及每笔交易的平均消费金额。根据每周交易日期数据，关于天使巧克力的交易量你能得出怎样的观察结果？请讨论。

3. 对每种浏览器类型汇总频数、消费总额以及每笔交易的平均消费金额。根据浏览器类型数据，关于天使巧克力的交易量你能得出怎样的观察结果？请讨论。

4. 绘制散点图并计算样本相关系数以探索在网站上度过时间和消费金额之间的关系，以在网站上度过时间为横轴。请讨论。

5. 绘制散点图并计算样本相关系数以探索观看网页数量和消费金额之间的关系，以观看网页数量为横轴。请讨论。

6. 绘制散点图并计算样本相关系数以探索在网站上度过时间和观看网页数量之间的关系，以观看网页数量为横轴。请讨论。

案例 3-5 非洲象数量

虽然数以百万计的大象曾经漫步于非洲，但是在20世纪80年代中期，非洲国家大象的数量因偷猎受到毁灭性打击。大象是非洲生态系统的重要组成部分。在热带森林，大象在有利于新树木生长的树冠中创造林中空地。在热带稀树草原，大象减少灌木覆盖，创造有利于食草动物的环境。另外，许多植物的种子在发芽前依赖于经过大象的消化道。

现在，在非洲大象的状况有很大不同；在一些国家，采取强有力的措施有效地保护了大象的数量，而与此同时，由于为了肉类和象牙的偷猎、栖息地的丢失以及人类的战争，在另一些国家大象继续处于危险之中。表3-13给出了1979年、1989年和2007年几个非洲国家的大象数量［Lemieux and Clarke，“The International Ban on Ivory Sales and Its Effects on Elephant Poaching in Africa，” British Journal of Criminology，49（4），2009］。

表3-13 1979年、1989年和2007年几个非洲国家的大象数量

国家	大象总数		
	1979	1989	2007
安哥拉	12 400	12 400	2 530
博茨瓦纳	20 000	51 000	175 487
喀麦隆	16 200	21 200	15 387
中非	63 000	19 000	3 334
乍得	15 000	3 100	6 435
刚果	10 800	70 000	22 102
刚果共和国	337 700	85 000	23 714
加蓬	13 400	76 000	70 637
肯尼亚	65 000	19 000	31 636
莫桑比克	54 800	18 600	26 088
索马里	24 300	6 000	70
苏丹	134 000	4 000	300
坦桑尼亚	316 300	80 000	167 003
赞比亚	150 000	41 000	29 231
津巴布韦	30 000	43 000	99 107

1977年成立的薛德瑞克野生动物基金会（David Sheldrick Wildlife Trust），是为了纪念自然学家戴维·莱斯利·威廉·薛德瑞克（David Leslie William Sheldrick），他创办了肯尼亚东查沃国家公园（Tsavo East National Park），领导该国的野生动物保护和管理规划单位。薛德瑞克基金会管理部门想知道，这些数据能显示出自1979年以来非洲各国大象总数的何种变动。

管理报告

使用描述统计学的方法汇总数据，并对自1979年以来非洲各国大象数量的变动进行评论。你的报告至少应该包括下面的内容：

1. 从1979~1989年，每个国家大象数据的年平均变化，并讨论在这10年中哪个国家大象的数量变化最大。
2. 从1989~2007年，每个国家大象数据的年平均变化，并讨论在这18年中哪个国家大象的数量变化最大。
3. 比较1和2中的结果，并讨论你从这些比较中得出的结论。

第4章 概率

CHAPTER 4

实践中的统计

概率在救援中的应用[一]

2010年8月5日，智利圣何塞铜矿发生塌方事故，33名矿工被埋在2000英尺[二]深的地下。圣何塞铜矿位于智利科比亚波市附近的阿塔卡马沙漠中。矿工们生存的希望渺茫，在焦虑中人们尝试用各种定位方法确定矿工所处的位置，以及他们是否仍然活着。事故发生后的第17天，营救人员借助5.5英尺粗的探杆发现了矿工并确认他们都还活着。

当务之急是尽快将这些矿工安全地带回地面，更为至关重要的是，营救工作必须谨慎进行。亨利·拉斯是承包商Murray&Roberts的经理，他在接受《圣地亚哥时报》的采访时说："由于圣何塞矿已开发多年，营救中要当心发生再次垮塌。"因此，营救方案必须周密设计，谨慎实施。

智利政府请求美国宇航局（NASA）充当营救行动的顾问。于是，NASA派出了一个由一名工程师（克林顿·克雷格）、两名物理学家（迈克尔·邓肯、J. D. 波尔克）以及一名心理学家（Al Holland）组成的四人小组。当被问及宇航局为何受邀做矿难救援的顾问时，邓肯说："在营救设备的设计以及长期独处所遇到的问题方面，我们可以向智利同行分享我们的经验。"

正如人们所想象的那样，失败的可能性还是相当高的。克雷格说："鉴于营救设备要在通道中穿行40次以上，因此必须考虑可能发生的零件磨损。我们建议预备三套营救设备并备好所需零件。为了提高救援成功的可能性，还要在笼子的四面安装带有弹簧加载的辊，使得笼子在营救通道中穿行时不会与周围的岩壁发生直接接触，防止笼子受到损毁。"

仔细考虑各种事件发生的概率尤为重要，但是我们却并不能精确地估计这些概率，只能由美国宇航局的科学家基于太空飞行中类似情况下的经验进行主观估计。邓肯说："虽然我们和智利方面的营救人员都希望根据历史数据能够精确地估计这些概率，但是矿难的唯一性却使得精确估计无法进行。"比如，在上升到地面的过程中，矿工必须垂直站立在笼子中，初步估计这大概需要历时2~4个小时。因此，必须考虑到昏迷。我们所能做的是，充分考虑营救中可能发生的各种情况，并运用航天员在执行短期和长期空间任务后返回地球过程中所积累的经验，对这些情况予以处置。邓肯最后总结说："实际上，将笼子从井底升到地面只需要15分钟。我们最初的估计是相当保守的。仔细考虑可能遇到的风险，这正是我们要做的。"

最终，智利经与美国宇航局商讨后设计的营救方案取得了圆满成功。10月13日，最后一名矿工出现；不锈钢制成的Fénix2救援舱长13英尺，重达924磅，承受住了在矿井中超过40次的往返穿行；在升回到地面的过程中，并没有矿工发生昏厥。

美国宇航局经常要在上述这种孤立情形下运用主观概率。空间飞行器上的空间有限，NASA必须估算各种零部件失灵的概率以及与之相伴的风险，从而决定在太空飞行中哪些零部件需要携带备用件。NASA还利用概率评价航天员在执行空间探险任务时的健康情况和工作绩效。

管理者经常是在分析不确定性的基础上进行决策的，比如：

（1）如果提高价格，则销售量下降的"可能性"有多大?

（2）某种新的装配方法提高生产效率的"可能性"有多大?

（3）某项工程如期完成的"可能性"有多大?

（4）某项新投资赢利的"可能性"有多大?

概率[三]（probability）是对某一特定事件发生的可能性的数值度量。因此，概率可以用来度量上述4个事件的不确定性程度。我们利用概率可以度量某个事件发生的可能性大小。

[一] 作者在此感谢迈克尔·邓肯博士和美国宇航局的克林顿·克雷格，他们为实践中的统计提供了本案例。

[二] 1英尺=0.3米。

[三] 对概率的研究最早起源于17世纪50年代Pierre de Fermat与Blaise Pascal之间的一系列通信。

事件的概率总是在0~1之间。若事件的概率等于0，则表明事件几乎不可能发生；若事件的概率等于1，则表明事件几乎肯定要发生；若事件的概率介于0和1之间，则代表事件发生可能性的程度。比如说，对于“明天下雨”这一事件，如果气象预报称“降水概率几乎为零”，则我们理解为明天几乎不可能下雨；如果天气预报称“降水概率为90%”，则意味着极有可能要下雨；如果天气预报称“降水概率为50%”，则有无降水的可能性是均等的。图4-1形象地描述了概率是如何用数值来度量事件发生可能性的大小的。

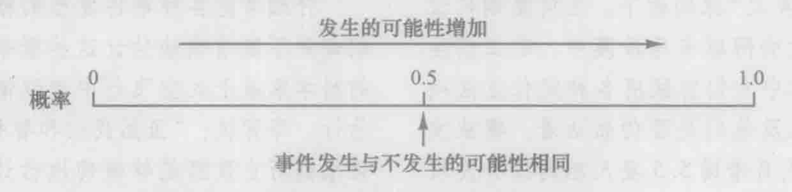


图4-1 概率是对事件发生可能性的数值度量

4.1 试验、计数法则和概率分配

在概率论中，我们把可以产生明确结果的过程定义为**试验**（experiment）。在一次试验中，有且只有一个试验结果发生。下面是几个有关试验及试验结果的例子。

试验	试验结果	试验	试验结果
抛掷一枚硬币	正面、反面	掷一枚色子	1，2，3，4，5，6
检测一个零件	合格、不合格	进行一场足球比赛	获胜、失利、平局
进行一次电话销售	购买、不购买		

一旦确定了试验的所有可能结果，就确定了试验的**样本空间**（sample space）。

样本空间

试验的样本空间是试验所有可能结果组成的一个集合。

任何一个特定的试验结果被称为**样本点**（sample point），它是样本空间中的一个元素。

上表中的第一个试验里，抛掷一枚硬币的试验结果（样本点）只有两种：硬币朝上的一面要么是正面要么是反面。用 S 表示样本空间，可以用下面的记号来表示样本空间：

$$S = \{正面,反面\}$$

对表中的第二个试验——检测一个零件，其样本空间为：

$$S = \{合格,不合格\}$$

上述两个试验都只有两种最基本的试验结果（样本点）。但是，对表中的第四个试验——抛掷一枚色子，试验结果为色子朝上一面的点数。此时，样本空间有6个点，为：

$$S = \{1,2,3,4,5,6\}$$

4.1.1 计数法则、组合和排列

对试验结果（样本点）进行确认和计数是计算概率的基础。下面，我们讨论三条有用的计数法则。

多步骤试验（multiple-step experiment） 第一条计数法则适用于多步骤试验。考虑抛掷两枚硬币的试验，试验结果是两枚硬币朝上一面的图案。在这个试验中，有多少种可能的试验结果呢？掷两枚硬币的试验可以视为一个两步骤的试验：第1步是抛掷第一枚硬币，第2步是抛掷第二枚硬币。令 H 表示正面朝上，T 表示反面朝上，则（H，H）表示第一枚硬币是正面朝上，第二枚硬币也是正面朝上。试验的样本空间 S 为：

$$S = \{(H,H),(H,T),(T,H),(T,T)\}$$

该试验有四种可能的结果，可以轻而易举地列出所有的试验结果。

利用多步骤试验的计数法则，我们可以确定出所有可能试验结果的数目，而不必将这些试验结果一一枚举。

多步骤试验的计数法则

如果一个试验可以分为循序的 k 个步骤，在第1步中有 n_1 种试验结果，在第2步中有 n_2 种试验结果…以此类推。那么所有可能的试验结果的总数为 $n_1 \times n_2 \times \cdots \times n_k$。

仍考虑抛掷两枚硬币的试验，它可以视为一个两步骤的试验。抛掷第一枚硬币有两种（$n_1=2$）不同的结果，抛掷第二枚硬币也有两种（$n_2=2$）不同的结果，根据计数法则共有 $2\times2=4$ 种不同的试验结果，得到 $S=\{(H, H), (H, T), (T, H), (T, T)\}$。在一个抛掷6枚硬币的试验中，有 $2^6=64$ 种不同的试验结果。

树形图（tree diagram）[⊖]是一种帮助分析多步骤试验的图形。图4-2是抛掷两枚硬币试验的树形图。由左到右循序经过两个步骤：第1步抛掷第一枚硬币；第2步抛掷第二枚硬币。在每个步骤中，都有两种可能的结果——正面朝上或者反面朝上，对于第1步里的每一种结果，都有两个分枝与第2步相对应。树型图右侧的每一个节点对应着试验的一种结果。从树型图左侧节点通往右侧的每一条路径对应着试验的一种结果。

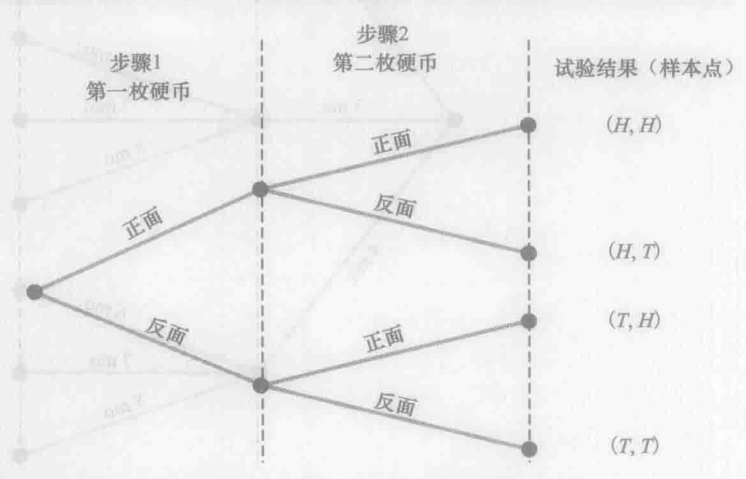


图4-2 抛掷两枚硬币的试验中的树形图

现在我们看看如何利用多步骤试验计数法则分析肯塔基光电（KP&L）公司的一个扩容工程。KP&L新建一项工程，旨在增加公司在肯塔基北部的一家工厂的发电能力。工程分为两个连续的阶段或步骤进行：阶段1（设计阶段）和阶段2（建设阶段）。虽然每个阶段都制定了尽可能详细的规划，但是管理人员仍然无法确切预知该项工程各个阶段完工的精确时间。对同类建筑工程的分析显示，设计阶段可能需要2、3或4个月，而建设阶段完工可能需要6、7或8个月。此外，由于对新增电力的急迫需求，管理人员计划整个工程在10个月内完工。

由于项目在设计阶段（步骤1）所需时间有3种可能，建设阶段（步骤2）所需时间也有3种可能，由多步骤试验计数法则可知，总共有 $3\times3=9$ 种试验结果。我们用一对数字来记录试验结果，比如，(2, 6) 表示设计阶段需要2个月完成，建设阶段需要6个月完成，即完成整个工程总共需要 $2+6=8$ 个月。在表4-1中汇总了KP&L问题中的9种试验结果。图4-3是树形图，由此可见这9种结果（样本点）是如何产生的。

计数法则和树形图能够帮助项目经理确定试验结果，并确定工程完工所需时间的概率。由图4-3可见，工程完工可能需要8至12个月，在9种结果中有6种能保证工程在10个月内完工。除了确定试验结果，我们还要考虑如何为各个试验结果分配概率值，以便最终评估工程在预期的10个月以内完工的可能性有多大。

表4-1 KP&L项目的试验结果（样本点）

完工时间（月）		试验结果	项目总的完工时间（月）	完工时间（月）		试验结果（样本点）	项目总的完工时间（月）
阶段1 设计	阶段2 建设			阶段1 设计	阶段2 建设		
2	6	(2, 6)	8	3	8	(3, 8)	11
2	7	(2, 7)	9	4	6	(4, 6)	10
2	8	(2, 8)	10	4	7	(4, 7)	11
3	6	(3, 6)	9	4	8	(4, 8)	12
3	7	(3, 7)	10				

⊖ 如果不利用树形图，我们可能认为抛掷2枚硬币总共有3种试验结果：0个正面、1个正面和2个正面。

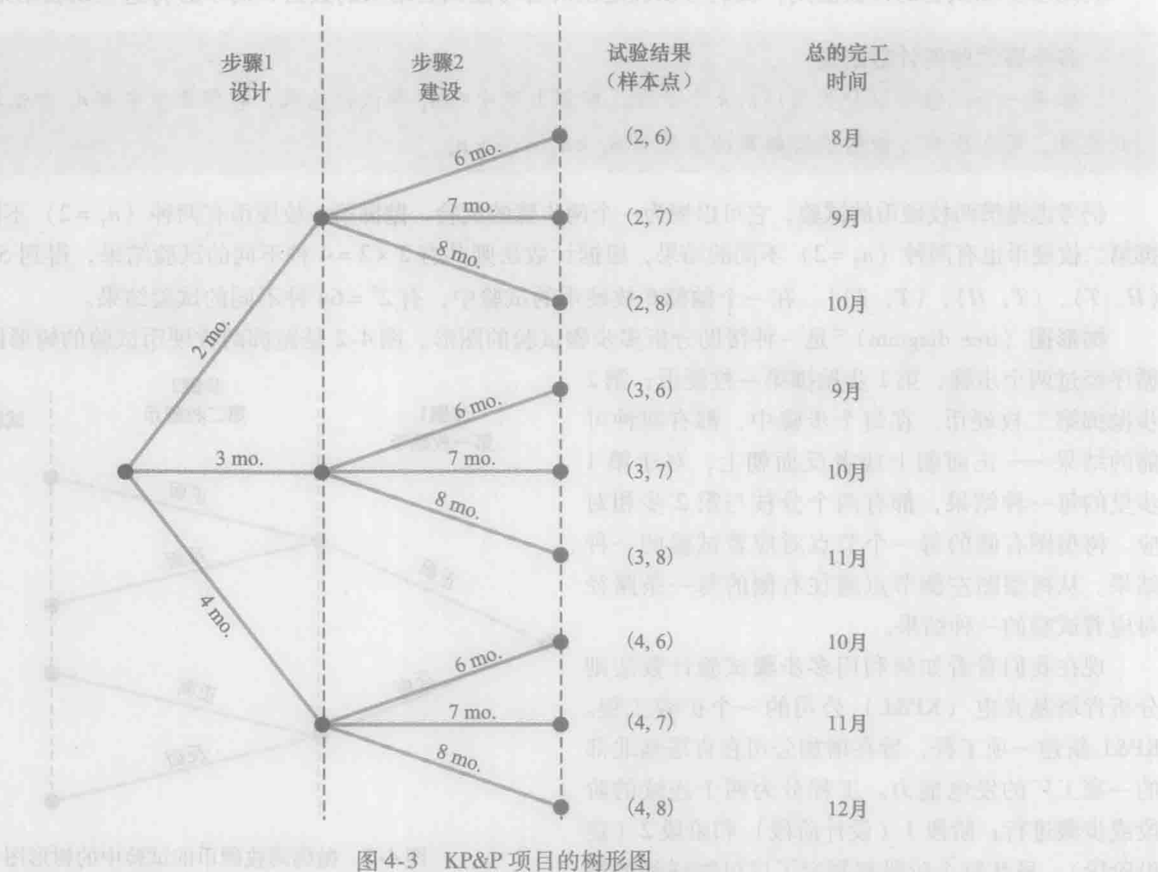


图 4-3 KP&P 项目的树形图

组合[⊖]（cornbinations） 在从 N（通常比较大）项中选取 n 项的试验中，还有另一种计数法则可以用于确定试验结果的数目，我们称这种计数法则为组合计数法则。

组合计数法则

从 N 项中任取 n 项的组合数为

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{4-1}$$

其中，$N! = N\times(N-1)\times(N-2)\times\cdots\times2\times1$

$n! = n\times(n-1)\times(n-2)\times\cdots\times2\times1$

并且定义

$$0! = 1$$

符号“!”表示阶乘。例如，5 的阶乘记作 $5! = 5\times4\times3\times2\times1 = 120$。

下面，以质检为例说明如何使用组合计数法则。假定质检人员要从 5 个零件中随机抽取 2 个零件进行检查，共有多少种不同选法呢？此时 $N=5$，$n=2$，由式（4-1）的组合计数法则，可得：

$$C_2^5 = \binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5\times4\times3\times2\times1}{(2\times1)(3\times2\times1)} = \frac{120}{12} = 10$$

因此，从一组 5 个零件中随机选取 2 个零件，共有 10 种可能的试验结果。如果我们把这 5 个零件分别标号为 A、

⊖ 从一个容量为 N 的总体中选取容量为 n 的样本时，可以利用组合计数法则计算样本的数目。

B、C、D 和 E，那么这 10 种组合或试验结果分别为 AB、AC、AD、AE、BC、BD、BE、CD、CE 和 DE。

再比如，佛罗里达州彩票发行系统每周从 53 个数码中随机抽取 6 个作为本周的中奖号码。利用式（4-1）的组合计数法则，可以确定从 53 个数中随机抽取 6 个数有多少种不同的方式。

$$\binom{53}{6} = \frac{53!}{6!(53-6)!} = \frac{53!}{6!47!} = \frac{53 \times 52 \times 51 \times 50 \times 49 \times 48}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = 22\ 957\ 480$$

组合计数法则说明，在彩票抽奖中大约有 2 300 多万种不同的试验结果[⊖]。购买一张彩票中奖的可能性仅为 1/22 957 480。

排列（permutalions）　第三种常用计数法则是排列，当从 N 项中选取 n 项并且考虑选取的顺序时（即对于选出的 n 项，不同的选择顺序被认为是不同的试验结果），排列可用于计算有多少种不同的试验结果。

排列计数法则

从 N 项中任取 n 项的排列数为

$$P_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \tag{4-2}$$

排列计数法则与组合计数法则具有密切的联系。由于任意选取的 n 项共有 $n!$ 种不同的排序方式，从而 P_n^N 总是大于 C_n^N。比如，在上述质量检验的例子中，质检人员从 5 个零件中随机抽取 2 个零件进行检查，则有多少种排列呢？此时 $N=5$，$n=2$，由排列计数法则式（4-2）可得：

$$P_2^5 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = \frac{120}{6} = 20$$

因此，在考虑顺序的情况下，从 5 个零件中随机抽取 2 个共有 20 种不同的结果。如果我们把这 5 个零件分别标号为 A、B、C、D 和 E，那么这 20 种排列分别为 AB、BA、AC、CA、AD、DA、AE、EA、BC、CB、BD、DB、BE、EB、CD、DC、CE、EC、DE 和 ED。

4.1.2 概率分配

现在我们来看一看如何为试验结果分配概率。古典法、相对频数法和主观概率法是最为常用的三种方法。无论采用何种方法，都必须满足**概率分配的两个基本条件**（basic requirements for assigning probabilities）。

概率分配的基本条件

1. 分配给每个试验结果的概率值都必须在 0 和 1 之间。若以 E_i 表示第 i 种试验结果，$P(E_i)$ 表示这种结果发生的概率，则有

$$0 \leqslant P(E_i) \leqslant 1, \quad（对所有的 i） \tag{4-3}$$

2. 所有试验结果的概率之和必须为 1。对 n 个试验结果的情形，有

$$P(E_1) + P(E_2) + \cdots + P(E_n) = 1 \tag{4-4}$$

古典法（classical method）当各种试验结果是等概率发生时，适合采用古典法进行概率分配。如果某一试验有 n 个可能的试验结果，则为每个试验结果分配的概率均为 $1/n$。这种方法下，概率分配的两个基本条件都成立。

比如：在抛掷一枚均匀硬币的试验中，有两种试验结果（正面或反面朝上），并且这两种结果出现的可能性相等。因此，正面朝上的概率为 1/2。同理，反面朝上的概率也是 1/2。

再比如：在抛掷一枚色子的试验中，6 种试验结果发生的可能性是均等的，因而为每种试验结果分配的概率都是 1/6。若以 $P(1)$ 表示朝上的一面为 1 点的概率，则 $P(1)=1/6$。同理，$P(2)=1/6$，$P(3)=1/6$，$P(4)=1/6$，$P(5)=1/6$，$P(6)=1/6$。可见，上述概率都满足式（4-3）和式（4-4）这两个概率分配的基本条件——每个概

⊖ 由组合计数法则可知，这类彩票中奖机会微乎其微。

率值都大于或者等于0，并且这些概率之和为1。

相对频数法（relative frequency method）相对频数法适用于试验可以大量重复进行，并且能取得试验结果发生比率的数据。比如，研究某地方医院X光检查的候诊人数，记录每天上午9点的候诊人数，得到以下连续20天的记录。

候诊人数	天数
0	2
1	5
2	6
3	4
4	3
合计	20

以上数据显示，在20天中有2天没有患者候诊，有5天有1名患者候诊，等等。根据相对频率法，为"没有患者候诊"分配概率2/20 =0.1，为"有1名患者候诊"分配概率5/20 =0.25，为"有2名患者候诊"分配概率6/20 =0.3，为"有3名患者候诊"分配概率4/20 =0.2，为"有4名患者候诊"分配概率3/20 =0.15。可见，相对频数法与古典法一样，也满足式（4-3）和式（4-4）这两个概率分配的基本条件。

主观法（subjective method）当不能假定试验结果是等可能发生或者无法取得相关数据时，可以采用主观法为试验结果分配概率。在采用主观法为试验结果分配概率时，可以利用一切可以获得的信息，比如我们的经验和直觉。当考虑所有可获得的信息后，主观法为试验结果分配一个（介于0和1之间的）概率值，用以表达对各个试验结果发生的可信程度。由于主观法表达的是某个人的置信度，它会因人而异：在主观法中，不同的人会对同一试验结果给出不同的概率值。

主观法要求特别注意保证式（4-3）和式（4-4）这两个基本条件成立。无论某人的可信程度有多大，对每一试验结果分配的概率值都必须在0与1之间，并且所有试验结果的概率值之和必须等于1。

例如，汤姆·埃尔斯本德和朱迪·埃尔斯本德申请购买一所住房，有两种可能的结果：

$$E_1 = \text{申请获得批准}$$

$$E_2 = \text{申请被驳回}$$

朱迪认为申请获得批准的可能性为0.8，因此她令 $P(E_1)=0.8$，$P(E_2)=0.2$。但汤姆认为申请获得批准的可能性为0.6，从而他令 $P(E_1)=0.6$，$P(E_2)=0.4$。汤姆对 E_1 概率的估计反映了他对申请获准持更消极的态度。

汤姆和朱迪所估计的主观概率都满足式（4-3）和式（4-4）这两个基本条件。虽然他们二人给出了不同的概率估计，但这恰恰体现了主观概率法的个性化特征。

在商务应用中，即便在古典法或者相对频数法适用的情况下，管理人员可能仍希望提供主观概率估计。此时，将古典法或相对频数法的估计与主观概率估计相结合，可得到相应概率的最佳估计。[⊖]

4.1.3 肯塔基光电公司（KP&L）项目的概率

为了深入分析KP&L公司项目，我们必须为表4-1中9种试验结果分配概率。管理层基于经验和判断，认为各种试验结果并非等可能发生。因此，不能使用古典法分配概率。于是，管理层决定对该公司在最近三年中同类工程的完工时间进行研究，表4-2中汇总了对40项类似项目的研究结果。

在浏览了研究结果以后，管理层决定采用相对频数法分配概率。本来，管理层还可以使用主观法分配概率，但是他们觉得当前的工程与以前进行的40项工程十分相似，因此认为相对频数法更优。

在利用表4-2中的数据计算概率时，试验结果（2，6）表示项目在阶段1用2个月和阶段2用6个月完工，我们注意到在40个项目中（2，6）发生了6次。根据相对频数法，为这一试验结果分配概率6/40 =0.15。类似地，试验结果（2，7）在40项工程中也发生了6次，它的概率即为6/40 =0.15。类似地，我们可以为表4-3中

⊖ 贝叶斯定理（见第4.5节）将主观确定的先验概率与其他方法得到的概率相结合，得到修正的后验概率。

KP&L公司项目的每个样本点分配概率。注意：符号 $P(2, 6)$ 表示样本点（2，6）的概率，$P(2, 7)$ 表示样本点（2，7）的概率，…，以此类推。

表4-2 40项KP&L公司工程的完工结果

完工时间（月）		样本点	以往工程数目
阶段1设计	阶段2建设		
2	6	(2，6)	6
2	7	(2，7)	6
2	8	(2，8)	2
3	6	(3，6)	4
3	7	(3，7)	8
3	8	(3，8)	2
4	6	(4，6)	2
4	7	(4，7)	4
4	8	(4，8)	6
		合计	40

表4-3 利用相对频数法为KP&L公司项目分配概率

样本点	工程完工时间（月）	样本点的概率
(2，6)	8	$P(2, 6) = 6/40 = 0.15$
(2，7)	9	$P(2, 7) = 6/40 = 0.15$
(2，8)	10	$P(2, 8) = 2/40 = 0.05$
(3，6)	9	$P(3, 6) = 4/40 = 0.10$
(3，7)	10	$P(3, 7) = 8/40 = 0.20$
(3，8)	11	$P(3, 8) = 2/40 = 0.05$
(4，6)	10	$P(4, 6) = 2/40 = 0.05$
(4，7)	11	$P(4, 7) = 4/40 = 0.10$
(4，8)	12	$P(4, 8) = 6/40 = 0.15$
		合计 1.00

注释和评论

1. 在统计学中，试验的概念与自然科学中试验的概念稍有不同。在自然科学中，试验通常是在试验室或者人工控制的环境中进行，目的是研究因果关系和效果。在统计学试验中，结果由概率决定。即使试验在完全相同的条件下重复进行，也可能会获得完全不同的结果。由于概率对试验结果的影响，有时也将统计学上的试验称为随机试验。
2. 在从容量为 N 的总体中无放回随机抽取一个容量为 n 的样本时，应该采用组合计数规则来计算有多少个不同的样本。

练习

方法

☆2. 6个项目组成一组，从中任取3项，共有多少种取法？以字母 A、B、C、D、E 和 F 分别表示这6项，请列出所有可能的组合。

4. 考虑将一枚硬币抛掷3次的试验。
   a. 画出试验的树形图。
   b. 列出试验的各种结果。
   c. 计算每种试验结果的概率。

☆6. 某项试验有3种结果：E_1、E_2 和 E_3，现将该试验重复50次，其中 E_1 出现20次，E_2 出现13次，E_3 出现17次。请为每种试验结果分配概率。你采用哪种方法分配概率？

应用

8. 在米尔福德市，申请改变城市区划需要经过计划委员会审议和市政委员会最终决定这两个步骤后方能实现。在第1阶段，计划委员会审议改变城市区划的申请，并给出支持或反对的意见。在第2阶段，市政委员会对计划委员会的意见进行投票，批准或

否决改变城市区划。假定某房地产开发商递交了一份申请，要求改变城市区划。

a. 若将申请程序视为一个试验，则在这个试验中有多少个样本点？请列出这些样本点。

b. 画出试验的树形图。

☆10. 许多学生在大学毕业前都有负债。下表中是来自4所大学和4所人文学院毕业生的负债比率和负债金额（《美国新闻和世界报道》《美国最好的大学》，2008年）。

大学	负债率（%）	负债金额（美元）
佩斯	72	32 980
爱荷华州立	69	32 130
马萨诸塞	55	11 227
纽约州立大学阿尔尼分校	64	11 856
学院	**负债率（%）**	**负债金额（美元）**
瓦尔特堡	83	28 758
莫尔豪斯	94	27 000
卫斯理女子学院	55	10 206
伍夫德	49	11 012

a. 如果随机选取一名莫尔豪斯学院的毕业生，则他（她）毕业时负债的概率是多少？

b. 在随后关于学生贷款的研究中，如果你从8所科研院校中随机选取一所，其学生毕业时的负债率超过60%的概率是多少？

c. 在随后关于学生贷款的研究中，如果你从8所科研院校中随机选取一所，则其学生毕业时负债金额的均值高于30 000美元的概率是多少？

d. 一名佩斯大学的毕业生没有负债的概率是多少？

e. 在佩斯大学毕业时负债的学生中，负债额的均值为32 980美元。佩斯大学的所有毕业生中，毕业生的人均负债是多少？

12. 在美国的28个州及维尔金群岛和哥伦比亚地区，强力球六合彩每周摇奖两次。彩票的规则如下：投注者首先购买一张彩票，然后从1~55个数字中选5个数，再从1~42个数字中选1个数作为强力球号码。每次摇奖时，彩票官员从装有55个白球的圆桶中抽出5个白球，再从装有42个红球的圆桶中抽出1个红球。要想获得头等奖，投注者所选数字必须与抽出的五个白球上的数字相同（顺序可以不同）且所选强力球号码与抽出的红球上的数字相同。2006年2月18日，内布拉斯加州林肯市康尼格拉食品公司的8名员工获得头等奖，奖金共计3.65亿美元，他们的中奖号码为15-17-43-44-49，强力球号码为29。除去头等奖之外，每期还有其他形式的现金奖。比如，如果所选数字与抽出的5个白球上的号码一致，则能够获得200 000美元的奖金（powerball网站，2006年3月19日）。

a. 抽出的5个白球上的号码有多少种不同的方式？

b. 与抽出5个白球上的号码一致，获得200 000美元奖金的概率是多少？

c. 获得强力球六合彩头等奖的概率是多少？

4.2 事件及其概率

在本章绪论中，我们对术语"事件"的运用仍限于日常用语中的含义。在第4.1节中，我们给出了有关试验以及与之相联系的试验结果（样本点）的概念。样本点和事件是研究概率论的基础。因此，下面必须给出与样本点相对应的**事件**（event）一个正式的定义。这是确定事件的概率的基础。

事件

事件是样本点的一个集合。

比如，在KP&L公司项目的问题中，假设项目经理所关心的是事件"整个工程可以在10个月以内完工"。由表4-3可见，有6个样本点——(2, 6)，(2, 7)，(2, 8)，(3, 6)，(3, 7) 和 (4, 6) ——可以保证工程在10个月以内完工。令C表示事件"工程可以在10个月以内完工"，则

$$C = \{(2,6),(2,7),(2,8),(3,6),(3,7),(4,6)\}$$

如果上述6个样本点中的任何一种试验结果出现，则称事件C发生。

KP&L公司管理层所关心的事件还包括：

L = 事件"工程完工时间少于10个月"

M = 事件"工程完工时间多于10个月"

运用表4-3的信息，可见上述事件含有以下样本点：

$$L = \{(2,6),(2,7),(3,6)\} \qquad M = \{(3,8),(4,7),(4,8)\}$$

对于KP&L公司的项目，还可以定义许多其他事件，但对每一事件来说，它都是试验中若干个样本点的集合。

根据表4-3列出的各样本点的概率，我们可以利用下面的定义计算KP&L公司管理层所关心的任一事件的概率。

事件的概率

事件的概率等于事件中所有样本点的概率之和。

利用这一定义，我们可以将某个特定事件中所有样本点（试验结果）的概率相加来计算事件的概率。现在，我们来计算“工程可以在10个月以内完工”这一事件的概率。由于事件 $C = \{(2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6)\}$，令 $P(C)$ 表示事件 C 的概率，有

$$P(C) = P(2,6) + P(2,7) + P(2,8) + P(3,6) + P(3,7) + P(4,6)$$

根据表4-3中样本点的概率可知

$$P(C) = 0.15 + 0.15 + 0.05 + 0.10 + 0.20 + 0.05 = 0.70$$

类似地，对于事件“工程完工时间少于10个月”，有 $L = \{(2, 6), (2, 7), (3, 6)\}$，从而事件的概率为

$$P(L) = P(2,6) + P(2,7) + P(3,6) = 0.15 + 0.15 + 0.10 = 0.40$$

最后，对于事件“工程完工时间多于10个月”，有 $M = \{(3, 8), (4, 7), (4, 8)\}$，因此

$$P(M) = P(3,8) + P(4,7) + P(4,8) = 0.05 + 0.10 + 0.15 = 0.30$$

根据上述结果可知，*KP&L* 公司在10个月或10个月以内完成工程的概率是0.70，完工时间少于10个月的概率是0.40，而完工时间多于10个月的概率是0.30。对于公司管理层所关心的其他事件，可以重复上述计算过程求得事件的概率。

只要我们能够确认一个试验的所有样本点并且为其分配概率，我们就能够根据定义来计算某一事件的概率。但是，若试验中存在大量的样本点，则会使得确认样本点和分配概率这一工作变得虽然不是不可能但却相当繁重。在本章余下的各节中，我们将介绍概率的一些基本性质，根据这些性质可以在不必知道每个样本点概率的情况下计算事件的概率。

注释和评论

1. 样本空间 S 是一个事件。由于它包含了所有的试验结果，它的概率为1，即 $P(S) = 1$。
2. 应用古典法分配概率时，假设试验结果是等可能发生的。在这种情况下，某一事件的概率等于这一事件所包含试验结果的数目除以试验结果总数。

练 习

方法

14. 某项试验有4种等可能发生的结果：E_1、E_2、E_3和E_4。求：
    a. E_2发生的概率是多少？
    b. 任意两种试验结果（比如：E_1或E_2）发生的概率是多少？
    c. 任意三种试验结果（比如：E_1或E_2或E_4）发生的概率是多少？

16. 在投掷一对色子的试验中，假定关注的是两枚色子朝上一面的点数之和。
    a. 共有多少种可能的样本点？（提示：利用多步骤试验计数法则。）
    b. 列出样本点。
    c. “两数之和为7”的概率是多少？
    d. “两数之和大于或者等于9”的概率是多少？
    e. 由于所有和数中包括6个偶数（2，4，6，8，10和12）和5个奇数（3，5，7，9和11），从而偶数出现的可能性比奇数更大。你是否同意上述观点？试解释原因。
    f. 你使用什么方法来分配概率？

应用

18. 为调查家庭在家就餐情况，哈里斯互动对496名与18周岁以下孩子共同生活的成人进行调查（《今日美国》，2007年1月3日）。调查结果见下表。

每周在家就餐的次数	调查中受访者人数	每周在家就餐的次数	调查中受访者人数
0	11	4	36
1	11	5	119
2	30	6	114
3	36	7次及以上	139

随机选择一个与18周岁以下孩子共同生活的家庭，试求：

a. 一周中未在家就餐的概率。

b. 一周中至少在家就餐4次的概率。

c. 一周中至多在家就餐2次的概率。

20.《财富》杂志公布了美国500强企业的年度名单。下表中数据是500强企业数目最多的五个州中500强企业的数目（*The New York Almanac*，2006年）。

州	公司数	州	公司数
纽约	54	伊利诺伊	33
加利福尼亚	52	俄亥俄	30
得克萨斯	48		

假定从500强企业中选取一家公司进行跟踪调查，求下列事件的概率。

a. 令 $N=\{$公司总部位于纽约州$\}$，求 $P(N)$。

b. 令 $T=\{$公司总部位于得克萨斯州$\}$，求 $P(T)$。

c. 令 $B=\{$公司总部位于以上五个州之一$\}$，求 $P(B)$。

4.3 概率的基本性质

4.3.1 事件的补

给定一个事件 A，定义**事件 A 的补**（complement of A）为“所有不包含在事件 A 中的样本点”，记为 A^c。图4-4称为**文氏图**（venn diagram），它用来说明补的概念。其中，矩形区域表示试验的样本空间，它包含了所有可能的样本点；圆形区域表示事件 A，它只包含了 A 中的样本点；矩形中的阴影部分是所有未包含在 A 中的样本点，记做 A^c。

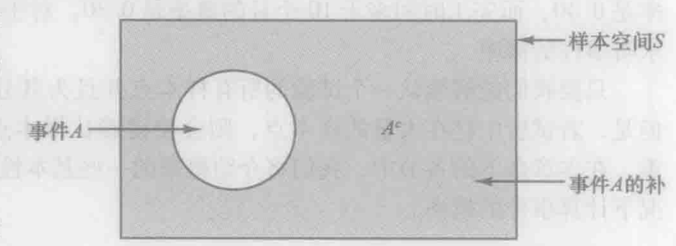


图4-4 事件 A 的补

在任何概率应用中，事件 A 和它的补 A^c 必有一个发生。因此，有

$$P(A)+P(A^c)=1$$

求解 $P(A)$，有如下结论。

利用事件的补计算概率

$$P(A)=1-P(A^c) \tag{4-5}$$

式（4-5）表明，如果已知事件 A 的补的概率 $P(A^c)$，就能容易地计算出事件 A 的概率。

比如，假定销售经理在看完销售报告后发现，新客户的合同中有80%并未成交。令事件 A 表示“合同成交”，A^c 来表示“合同没有成交”，该经理认为 $P(A^c)=0.8$，利用式（4-5），有

$$P(A)=1-P(A^c)=1-0.80=0.20$$

我们可以得出结论，认为新客户的合同中有20%成交。

再比如，假设某采购部门认为供货商运来的货物中无残次品的概率为90%，利用式（4-5），我们可以推断货物中有残次品的概率为 $1-0.90=0.10$。

4.3.2 加法公式

当求两个事件中至少有一个事件发生的概率时，加法公式非常有用。也就是说，对于两个事件 A 和 B，我们希望知道事件 A 或事件 B 或两者都发生的概率时，常常采用加法公式。

在讲述加法公式之前，我们需要首先讨论与事件关系有关的两个概念：事件的并和事件的交。对两个事件 A 和事件 B，定义**事件 A 和事件 B 的并**（union of A and B）如下：

两个事件的并

A 和 B 的并是所有的属于 A 或 B 或同时属于二者的样本点构成的事件，记作 $A \cup B$。

图 4-5 的文氏图描绘了事件 A 和 B 的并。注意：两个圆形区域包含了属于事件 A 和事件 B 的所有样本点。这两个圆形区域是有重叠的，这说明有一些样本点是既属于事件 A 又属于事件 B 的。

定义**事件 A 和事件 B 的交**（intersection of A and B）如下：

两个事件的交

给定两个事件 A 和 B，则 A 和 B 的交是同时属于 A 和 B 的样本点构成的事件，记作 $A \cap B$。

图 4-6 的文氏图描述了 A 和 B 的交。图中两个圆形区域相互重叠的部分即为两个事件的交，由同时属于 A 和 B 的样本点组成。



图 4-5 阴影部分是事件 A 和事件 B 的并　　图 4-6 阴影部分是事件 A 和事件 B 的交

下面继续介绍加法公式。**加法公式**（addition law）提供了一种计算"事件 A 发生或 B 发生或 A、B 同时发生"的概率的方法。也就是说，加法公式用来计算两个事件的并的概率。加法公式可以写成如下形式：

加法公式

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4-6}$$

从直观上看，加法公式右边的前两项 $P(A) + P(B)$ 已经包含了所有 $A \cup B$ 的样本点。但是由于 $A \cap B$ 内的样本点既属于 A 又属于 B，在计算 $P(A) + P(B)$ 时 $A \cap B$ 内的样本点被计算了两次。因此，需从中减去 $P(A \cap B)$，对重复计算进行修正。

作为一个加法公式的实际应用例子，我们考虑一个有 50 名工人的小型装配厂。每个工人都必须按时完成工作定额，并且使所装配的产品通过最终检验。有时，一些工人由于没能按时完成工作或者因为装配的产品不合格而没有达到工作标准。在期末的绩效评估中，生产经理发现 50 名工人中有 5 人未能按时完成工作，6 人装配的产品不合格，2 人既没能按时完成工作又存在装配不合格的问题。

令事件

$$L = \{未能按时完成工作\} \quad D = \{装配的产品不合格\}$$

根据相对频数法，有

$$P(L) = 5/50 = 0.10$$
$$P(D) = 6/50 = 0.12$$
$$P(L \cap D) = 2/50 = 0.04$$

生产经理在浏览了工作绩效数据后，决定在绩效评分中对所有未能按时完成工作或者装配的产品不合格的工人给予低分。经理所关心的事件为 $L \cup D$。生产经理为事件"绩效评分是低分"分配的概率是多少？

求两个事件的并的概率。尤其是，我们希望求出 $P(L \cup D)$。利用式（4-6），有

$$P(L \cup D) = P(L) + P(D) - P(L \cap D)$$

等式右边的三个概率值均为已知，可得

$$P(L \cup D) = 0.10 + 0.12 - 0.04 = 0.18$$

这一结果告诉我们，若随机选取一名工人则他的绩效评分是低分的概率为0.18。

考虑加法公式的另一个例子。某大型计算机软件公司人事部经理最近做了一项调查，他发现近两年内离职的公司雇员中有30%的人是因为对工资不满意，有20%的人是因为对分配的工作不满意，12%的人是因为对工资和分配的工作都不满意。那么在两年内离职的公司雇员中，其离职原因是因为对工资不满意、对分配的工作不满意，或者二者皆有的概率是多少呢？

令事件

$$S = \{公司雇员离职是因为对工资不满意\}$$
$$W = \{公司雇员离职是因为对分配的工作不满意\}$$

已知 $P(S)=0.30$，$P(W)=0.20$，$P(S \cap W)=0.12$。利用加法公式（4-6），可得

$$P(S \cup W) = P(S) + P(W) - P(S \cap W) = 0.30 + 0.20 - 0.12 = 0.38$$

得出雇员因为工资或工作原因离职的概率为0.38。

在结束加法公式的讨论之前，我们再来考虑一种特殊的情况：**互斥事件**（mutually exclusive events）。

互斥事件

如果两个事件没有公共的样本点，则称这两个事件互斥。

当事件 A 和事件 B 中一个发生而另一个一定不会发生时，则称 A 和 B 是互斥的。因此，A 和 B 互斥的必要条件是它们的交不含有任何样本点。文氏图描述了两个事件 A 和 B 互斥的情况，见图4-7。此时，$P(A \cap B)=0$，故加法公式可以写成下面的形式：

互斥事件的加法公式

$$P(A \cup B) = P(A) + P(B)$$

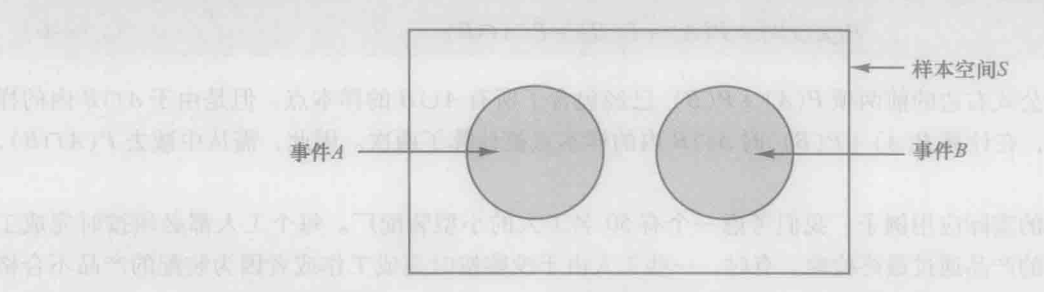


图4-7 互斥事件

练习

方法

22. 假设样本空间包含有5个等可能发生的试验结果：E_1、E_2、E_3、E_4和E_5。令

    $A = \{E_1, E_2\}$ $B = \{E_3, E_4\}$ $C = \{E_2, E_3, E_5\}$

    a. 求 $P(A)$、$P(B)$ 和 $P(C)$。

    b. 求 $P(A \cup B)$。A 和 B 是否互斥？

    c. 求 A^c、C^c、$P(A^c)$ 和 $P(C^c)$。

    d. 求 $A \cup B^c$ 和 $P(A \cup B^c)$。

    e. 求 $P(B \cup C)$。

应用

24. 克拉克森大学对毕业生进行了一项调查以了解他们对学校的看法。调查中的一个问题是询问受访者在学校的经历是未达到预期、达到预期还是超出预期？调查结果表明，有4%的人没有回答，有26%的人认为未达到预期，65%的人认为达到预期。

    a. 如果我们随机地选择一名毕业生，则他认为在学校的经历超出预期的概率为多少？

b. 如果我们随机地选择一名毕业生，则他认为在学校的经历达到或超出预期的概率为多少？

26. 晨星投资提供了互助基金的一些信息，内容包括：互助基金的类型（国内产权、国际产权或固定收益）和基金的晨星评级。评级从1星（最低级别）到5星（最高级别）。从晨星基金500(2008) 中选取25只基金组成一个样本，得到以下结果：
   - 有16只互助基金是国内产权基金
   - 有13只互助基金的评级是3星或者3星以下
   - 国内产权基金中有7只的评级是4星
   - 国内产权基金中有2只的评级是5星

   为了更多地了解互助基金及其投资目标，假设从这25只互助基金中随机选取一只基金。试问：

   a. 所选基金是国内产权基金的概率是多大？

   b. 所选基金的评级是4星或5星的概率是多大？

   c. 所选基金是国内产权基金并且评级是4星或5星的概率是多大？

   d. 所选基金是国内产权基金或者评级是4星或5星的概率是多大？

☆28. 某杂志对订阅者的调查表明，在过去12个月中，45.8%的人由于工作原因租赁汽车，54%的人由于个人原因租赁汽车，30%的人由于工作和个人原因租赁汽车。

   a. 在过去12个月中，某订阅者因工作或个人原因租赁汽车的概率是多少？

   b. 在过去12个月中，某订阅者由于工作或个人原因未租赁汽车的概率是多少？

4.4 条件概率

某个事件发生的可能性经常会受到另一个相关事件发生与否的影响。假设事件 A 发生的概率为 $P(A)$，如果获得了新的信息——确知另一个相关事件 B 已经发生了，我们希望利用这一新的信息来重新计算事件 A 发生的可能性。此时，事件 A 发生的可能性叫作**条件概率**（conditional probability），记作 $P(A\mid B)$。符号"|"用来表明我们是在事件 B 已经发生的条件下考虑 A 发生的可能性。因此，符号 $P(A\mid B)$ 读作"事件 B 发生的条件下事件 A 发生的概率"。

下面，以美国东部某大城市警察局男性和女性警官的升职情况为例，说明条件概率的应用。警察局共有1 200名警官，男性960人，女性240人。在过去的两年中，有324名警官得到了升职。其中男性和女性升职情况的详细数据见表4-4。

表4-4 过去两年中警官升职的情况

	男性	女性	合计
升职人数	288	36	324
未升职人数	672	204	876
合计	960	240	1 200

一个由女性警官组成的委员会在浏览了升职记录以后指出，在升职过程中存在性别歧视，其依据是男性警官中有288人得到了升职，而升职的女性警官仅为36人。警察局官员回应称，女性警官升职人数较少并非因为性别歧视，而是因为警官中女性的数量原本就相对较少。现在我们利用条件概率，对升职中性别歧视的投诉进行分析。

令

$M=\{$某警官为男性$\}$　　$W=\{$某警官为女性$\}$

$A=\{$某警官得到升职$\}$　　$A^c=\{$某警官未得到升职$\}$

将表4-4中各项数据除以警官总人数1 200，便得出以下有用的概率信息。

随机选择一名警官，他是男性并且得到了升职的概率为 $P(M\cap A)=288/1\,200=0.24$

随机选择一名警官，他是男性但未得到升职的概率为 $P(M\cap A^c)=672/1\,200=0.56$

随机选择一名警官，她是女性并且得到了升职的概率为 $P(W\cap A)=36/1\,200=0.03$

随机选择一名警官，她是女性但未得到升职的概率为 $P(W\cap A^c)=204/1\,200=0.17$

上述概率值都是两个事件交的概率，我们称之为**联合概率**（joint probability）。表4-5汇总了警官升职情况的概率信息，称为**联合概率表**。

表4-5 警官升职的联合概率分布表

联合概率

	男性(M)	女性(W)	合计
升职(A)	0.24	0.03	0.27
未升职(A^c)	0.56	0.17	0.73
合计	0.80	0.20	1.00

边际概率

在联合概率表的边缘分别列出了每个事件各自的概率，即 $P(M)=0.80$，$P(W)=0.20$，$P(A)=0.27$ 和 $P(A^c)=0.73$。因为这些概率位于联合概率表的边缘，故称之为**边际概率**（marginal probability）。我们注意到，边际概率总可以由联合概率表中的联合概率按行或按列求和得到。比如，得到升职的边际概率为 $P(A)=P(M\cap A)+P(W\cap A)=0.24+0.03=0.27$。由边际概率可知，80% 的警官为男性，20% 的警官为女性，27% 的警官得到了升职，73% 的警官未得到升职。

现在我们利用条件概率进行分析，计算某位警官是男性的条件下获得升职的概率。按条件概率的表示符号，即我们想要计算 $P(A\mid M)$。为此，首先我们认识这一符号的简单含义。$P(A\mid M)$ 度量的是在事件 M（某警官为男性）已经发生的条件下事件 A（得到升职）发生的概率。因此，我们需要关注的只是 960 名男性警官的升职情况。由于在 960 名男性警官中，有 288 名得到升职，则在某警官为男性的条件下得到升职的概率为 $288/960=0.30$。也就是说，若警官为男性，则他在过去两年之中获得升职的机会为 30%。

由于表 4-4 中的值已给出了每一类警官的人数，所以这种方法应用起来十分简单。现在我们演示怎样直接通过事件概率，而不是使用表 4-4 的频数数据计算条件概率 $P(A\mid M)$。

已知 $P(A\mid M)=288/960=0.30$，将分子分母同时除以警官总人数 1 200，有

$$P(A\mid M)=\frac{288}{960}=\frac{288/1200}{960/1200}=\frac{0.24}{0.80}=0.30$$

由此可见，条件概率 $P(A\mid M)$ 可以通过 $0.24/0.80$ 来计算。由联合概率表（见表 4-5）可知，事件 A 和 M 的联合概率 $P(A\cap M)=0.24$，事件"随机选择某警官为男性"的边际概率 $P(M)=0.80$。因此，条件概率 $P(A\mid M)$ 可以通过求联合概率 $P(A\cap M)$ 和边际概率 $P(M)$ 的比值得出，即：

$$P(A\mid M)=\frac{P(A\cap M)}{P(M)}=\frac{0.24}{0.80}=0.30$$

既然条件概率可以由联合概率和边际概率的比值计算得到，则对于任何两个事件 A 和 B，有如下条件概率的通用计算公式：

条件概率

$$P(A\mid B)=\frac{P(A\cap B)}{P(B)} \tag{4-7}$$

或

$$P(B\mid A)=\frac{P(A\cap B)}{P(A)} \tag{4-8}$$

文氏图可以帮助我们对条件概率有一个直观的理解。图 4-8 右边的圆表示事件 B，其中与事件 A 重叠的部分表示事件 $(A\cap B)$。可见，一旦事件 B 已经发生，则能够观测到事件 A 发生的唯一区域是事件 $(A\cap B)$。于是，$P(A\cap B)/P(B)$ 给出了在事件 B 已经发生的情况下事件 A 发生的条件概率。

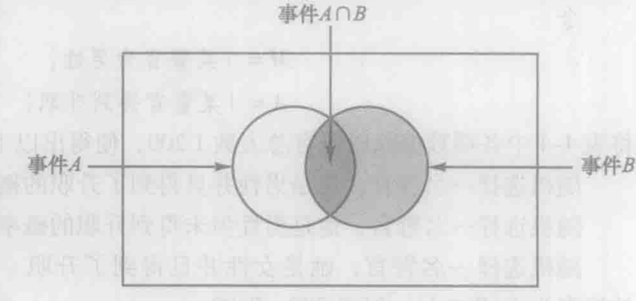


图 4-8 条件概率 $P(A\mid B)=P(A\cap B)/P(B)$

现在回到有关女警官升职中受到歧视的问题中来。表 4-5 中第一行的边际概率显示，某警官升职的概率是 $P(A)=0.27$（无论该警官是男是女）。在这一性别歧视的案例中，两个条件概率 $P(A\mid M)$ 和 $P(A\mid W)$ 是问题的关键。也就是说，若警官为男性，则他的升职概率为多少？若警官为女性，则她的升职概率为多少？如果这两个概率值相等，那么有关性别歧视的投诉就是毫无根据的，因为对于男女警官，有着均等的升职机会。但是，如果这两个条件概率不相等，则有证据表明男女警官在升职问题上被区别对待。

已知 $P(A\mid M)=0.30$，现在我们利用表 4-5 中的概率值和条件概率基本关系式（4-7），在某位警官为女性的

条件下计算其升职的概率 $P(A \mid W)$。利用式（4-7），其中 $B=W$，我们得到

$$P(A \mid W) = \frac{P(A \cap W)}{P(W)} = \frac{0.03}{0.20} = 0.15$$

你由此得到了什么结论？男性警官升职概率为0.30，是女性警官升职概率0.15的两倍。虽然条件概率本身并不能证明存在歧视，但条件概率的数值却成为支持女警官们投诉的有力证据。

4.4.1 独立事件

在上述例子中，$P(A)=0.27$，$P(A \mid M)=0.30$，并且 $P(A \mid W)=0.15$。我们看到，警官的性别大大影响了获得升职（事件 A）的概率。特别地，由于 $P(A \mid M) \neq P(A)$，我们可以说事件 A 和 M 是相依事件，即事件 A（升职）的概率受到事件 M（警官是男性）发生与否的影响。类似地，因为 $P(A \mid W) \neq P(A)$，我们可以说事件 A 和 W 是相依事件。如果事件 A 的概率不会由于事件 M 的存在而改变，即 $P(A \mid M)=P(A)$，则称事件 A 和 M 是**独立事件**（independent events）。由此，得到如下关于两个事件独立的定义：

独立事件

两个事件 A 和 B 是相互独立的，如果

$$P(A \mid B) = P(A) \tag{4-9}$$

或

$$P(B \mid A) = P(B) \tag{4-10}$$

否则，两个事件是相依的。

4.4.2 乘法公式

加法法则用来计算两事件并的概率，而乘法公式则是用来计算两事件交的概率。乘法公式是以条件概率的定义为基础的。利用式（4-7）和式（4-8）求解 $P(A \cap B)$，便可得到**乘法公式**（multiplication law）。

乘法公式

$$P(A \cap B) = P(B)P(A \mid B) \tag{4-11}$$

或

$$P(A \cap B) = P(A)P(B \mid A) \tag{4-12}$$

下面，我们举例说明如何应用乘法公式。假设某报纸的发行部已知在某社区有84%的住户订阅了该报纸的日报。令 D 表示事件"某住户订阅了日报"，则 $P(D)=0.84$。此外，还知道某个已经订阅日报的住户订阅周日特刊（事件 S）的概率为0.75，即 $P(S \mid D)=0.75$。那么，某住户既订阅日报又订阅周日特刊的概率为多少？利用乘法公式，有：

$$P(S \cap D) = P(D)P(S \mid D) = 0.84 \times 0.75 = 0.63$$

可见，有63%的住户既订阅了日报又订阅了周日特刊。

在结束本节之前，我们再来看一下当事件是相互独立这一特殊情况下乘法公式的形式。当 $P(A \mid B)=P(A)$ 或者 $P(B \mid A)=P(B)$ 时，事件 A 和 B 独立。因此，对两事件独立的特殊情况，利用式（4-11）和式（4-12）得到如下的乘法公式。

独立事件的乘法公式

$$P(A \cap B) = P(A)P(B) \tag{4-13}$$

为了计算两独立事件的交的概率，我们只需简单地将它们相应的概率相乘。因此，独立事件的乘法公式提供了另外一种判断 A 和 B 是否独立的方法，即如果 $P(A \cap B)=P(B)P(A)$，则 A 和 B 是独立的；如果 $P(A \cap B) \neq P(B)P(A)$，则 A 和 B 是相依的。

下面是独立事件的乘法公式的一个应用实例。假设根据以往经验一名加油站的经理已知有80%的顾客加油时使用信用卡。那么两名顾客加油时都使用信用卡的概率为多少？令

$$A=\{第一名顾客使用信用卡\} \quad B=\{第二名顾客使用信用卡\}$$

我们感兴趣的是两事件的交$A\cap B$。在并无其他信息的情况下，我们有理由假设事件A和事件B是独立的，因此，

$$P(A\cap B)=P(A)P(B)=0.80\times0.80=0.64$$

总的来说，我们研究条件概率的动机在于事件之间常常是相互关联的。在这种情形下，由于事件是相依的，所以我们计算条件概率时必须采用式（4-7）和式（4-8）；如果两个事件无关，它们是独立的，则一个事件的概率就不会受到另外事件发生与否的影响。

注释和评论

不要混淆互斥事件和独立事件，这是两个不同的概念。两个概率不为零的事件不可能既是互斥事件，又是独立事件。如果两个互斥的事件之一被确知已经发生了，那么另一事件发生的概率降为零，故它们是相依的。

练习

方法

☆30. 假设对于事件A和B，有$P(A)=0.50$，$P(B)=0.60$和$P(A\cap B)=0.40$。

a. 求$P(A\mid B)$。

b. 求$P(B\mid A)$。

c. A和B是否独立？为什么？

应用

32. 2009年1月，美国汽车行业销售657 000辆汽车（《华尔街日报》，2009年2月4日）。由于经济状况持续衰退，自2008年1月以来汽车销量下降了37%。通用、福特和克莱斯勒是美国三家最大的汽车制造商，它们的销量为280 500辆，自2008年1月以来下降了48%。下表是对销售量按照汽车制造商和售出车型进行的汇总（单位：千辆）。其中，非美国汽车制造商包括丰田、本田和日产等。轻型卡车包括：皮卡、微型货车、SUV和混合车型。

		车型	
		小汽车	轻型卡车
制造商	美国	87.4	193.1
	非美国	228.5	148.0

a. 根据上述数据求联合概率分布表，并根据该表回答下列问题。

b. 求边际概率，分别计算制造商和售出车型的概率？

c. 如果一辆车是由美国汽车商制造，则这是一辆小汽车的概率为多少？这是一辆轻型卡车的概率为多少？

d. 如果一辆车不是由美国汽车商制造，则这是一辆小汽车的概率为多少？这是一辆轻型卡车的概率为多少？

e. 如果一辆车是轻型卡车，则它是由美国汽车商制造的概率为多少？

f. 根据以上概率信息，你对销量得出什么结论？

34. 据美国交通运输部报告，11月份西南航空航班的准点到达率为83.4%；美国航空航班的准点到达率为75.1%；蓝捷航空航班的准点到达率为70.1%（《今日美国》，2007年1月4日）。假定上述航班的准点到达率对罗切斯特国际机场A候机大楼也是适用的。并且，在A候机大楼抵港航班中有40%是西南航空公司的航班，有35%是美国航空公司的航班，有25%是蓝捷航空公司的航班。

a. 编制联合概率分布表，表由三行（三家航空公司）和两列（准点情况）组成。

b. 刚刚有通告称，航班1424将在A候机大楼20通道抵港。则航班1424最可能属于哪家航空公司？

c. 航班1424准点到达的概率有多大？

d. 假定通告称航班1424延误，则航班1424最可能属于哪家航空公司？最不可能属于哪家航空

公司？

36. 贾马尔·克劳福德是美国职业篮球联赛中波特兰开拓者队最好的罚球得分手，他投球命中率为93%（ESPN网站，2012年4月5日）。假设在随后的一场篮球赛中，由于对方犯规，贾马尔·克劳福德获得两次罚球机会。
    a. 他两次罚球都命中的概率是多少？
    b. 他至少有一次罚球命中的概率是多少？
    c. 他两次罚球都没有命中的概率是多少？
    d. 在篮球比赛的最后时刻，球队经常采取故意犯规的战术以拖延比赛。通常的战术是故意对对方罚球技术最差的球员犯规。假设波特兰开拓者队中锋的罚球命中率为58%，若中锋获得两次罚球机会，请分别计算中锋出现（a），（b）和（c）情形的概率，并说明为什么说对波特兰开拓者队的中锋犯规是比对贾马尔·克劳福德犯规更好的战术。

38. 纽约州要求3~8年级学生参加一项全州的数学考试。必须证明学生已达到与其所在年级相对应的数学理解能力后，方能被认为达到了州里的熟练标准。下表中数据给出的是纽约市学校系统3~8年级参加考试学生中达标和未达标学生的人数（纽约市教育局网站，2010年1月16日）。

年级	是否达到熟练掌握程度	
	达到	未达到
三年级	47401	23975
四年级	35020	34740
五年级	36062	33540
六年级	36361	32929
七年级	40945	29768
八年级	40720	31931

    a. 根据上述数据，建立联合概率分布表。
    b. 求边际概率，由此得到学生是否达到熟练掌握程度的概率分别是多少？
    c. 若随机选取一名三年级学生，则他（她）达到熟练掌握程度的概率是多少？
    d. 若随机选取的一名学生在考试中已经达到熟练掌握的程度，则他（她）是一名三年级学生的概率有多大？他（她）是一名四年级学生的概率有多大？

4.5 贝叶斯定理

在条件概率的讨论中，我们指出在获得新的信息之后对概率进行修正是重要的概率分析手段。通常，在开始分析时，总是对所关心的特定事件估计一个初始或**先验概率**（prior probability）。然后，当我们从样本、专项报告或产品检验中获取了有关该事件新的信息时，就能根据这些新增信息计算修正概率对先验概率值进行更新，得到**后验概率**（posterior probability）。**贝叶斯定理**（Bayes's theorem）提供了进行这种概率计算的一种方法。图4-9列示了概率修正过程中的步骤。

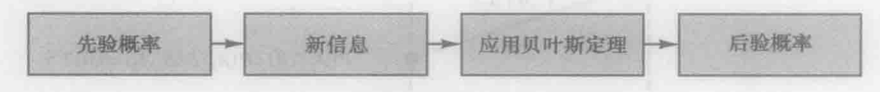


图4-9 利用贝叶斯定理进行概率修正

下面给出贝叶斯定理的一个应用实例。假设某制造企业从两个不同的供应商处购买零件。令A_1表示事件“零件来自供应商1”，A_2表示“零件来自供应商2”。

现在，该企业有65%的零件购自供应商1，其余35%的零件购自供应商2。那么，如果随机选取一个零件，则我们可设定先验概率$P(A_1)=0.65$和$P(A_2)=0.35$。

表4-6 两个供应商的零件质量的历史信息

	合格品率	不合格品率
供应商1	98	2
供应商2	95	5

零件的质量随货源的不同而不同，两个供应商零件质量的历史数据见表4-6中。如果令G表示事件“零件合格”，B表示事件“零件不合格”，则表4-6中的信息给出如下条件概率的值：

$$P(G\mid A_1)=0.98 \qquad P(B\mid A_1)=0.02$$
$$P(G\mid A_2)=0.95 \qquad P(B\mid A_2)=0.05$$

工厂首先从两个供应商之一购得零件，然后再检验某个零件质量是否合格。图4-10的树形图将这一过程描述为一

个两步骤试验——可见最终有4钟可能的试验结果，其中2个结果对应与质量合格，2个结果对应于零件质量不合格。

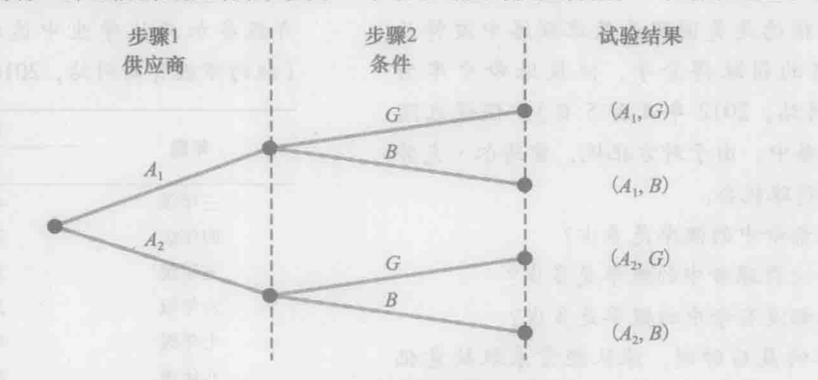


图4-10 两个供应商例子中的树形图

注：步骤1表明零件来自两个供应商之一。
步骤2表明零件质量是否合格。

每个试验结果都是两个事件的交，于是可以利用乘法公式来计算概率，比如：

$$P(A_1,G) = P(A_1 \cap G) = P(A_1)P(G \mid A_1)$$

计算联合概率的过程可借助概率树来描述，如图4-11所示。概率树从左到右有两个分支，步骤1的每一分枝的概率都是先验概率，步骤2的每一分支的概率都是条件概率。为了求得每一试验结果的概率，我们只需简单地把通向试验结果的各分支上的概率相乘即可。图4-11显示了每一分枝上的概率以及联合概率值。

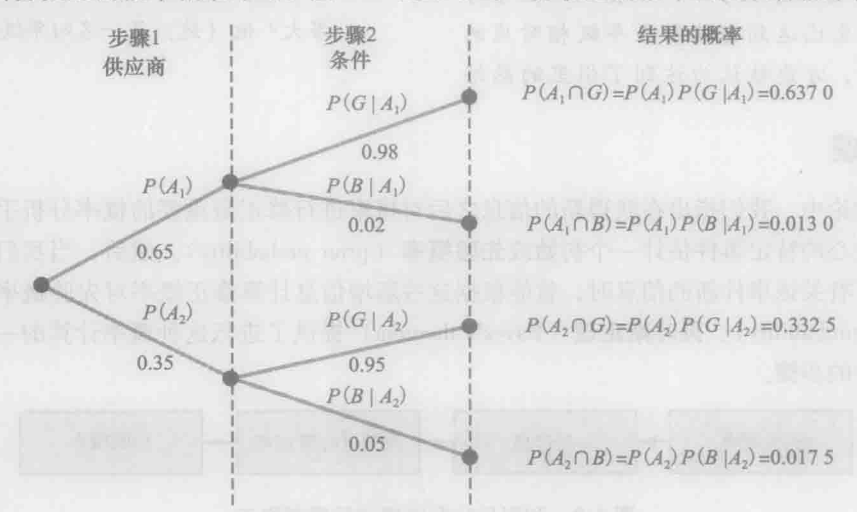


图4-11 两个供应商例子的概率树

现在假定购自两个供应商的零件都被用于加工程序中，一台机器因为遇到一个不合格零件而突然停机。那么在已知零件不合格的信息后，这个零件来自供应商1的概率为多少？这个零件购自供应商2的概率为多少？根据概率树上的信息（见图4-11），贝叶斯定理就可用于解答这样的问题。

令B表示“零件不合格”，我们现在想要求后验概率$P(A_1 \mid B)$和$P(A_2 \mid B)$。根据条件概率公式，有

$$P(A_1 \mid B) = \frac{P(A_1 \cap B)}{P(B)} \tag{4-14}$$

从概率树中可见

$$P(A_1 \cap B) = P(A_1)P(B \mid A_1) \tag{4-15}$$

由于事件B只在两种情况下才发生（$A_1 \cap B$和$A_2 \cap B$），因此有

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) = P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) \tag{4-16}$$

将式（4-15）和式（4-16）代入式（4-14）中，即可求得 $P(A_2 \mid B)$ 的值。在两个事件的情形下，得到如下的贝叶斯定理：

贝叶斯定理[⊖]（两事件的情形）

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \tag{4-17}$$

$$P(A_2 \mid B) = \frac{P(A_2)P(B \mid A_2)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \tag{4-18}$$

利用式（4-17）和本例中的概率值，有

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)}$$

$$= \frac{0.65 \times 0.02}{0.65 \times 0.02 + 0.35 \times 0.05} = \frac{0.0130}{0.0130 + 0.0175} = \frac{0.0130}{0.0305} = 0.4262$$

此外，利用式（4-18）可得 $P(A_2 \cap B)$：

$$P(A_2 \mid B) = \frac{0.35 \times 0.05}{0.65 \times 0.02 + 0.35 \times 0.05} = \frac{0.0175}{0.0130 + 0.0175} = \frac{0.0175}{0.0305} = 0.5738$$

在这个应用中，开始时随机选取的零件有0.65的概率来自供应商1。但是，在给定了零件不合格的信息以后，则这个零件来自供应商1的概率降低到0.4262。事实上，如果一个零件质量不合格，那么它就有超过一半的概率是来自于供应商2的，即 $P(A_2 \mid B) = 0.5738$。

贝叶斯定理通常应用于如下情况，即我们希望计算后验概率的那些事件是互斥的，且它们的并构成了整个样本空间[⊜]。对 n 个互斥事件 $A_1, A_2, \cdots, A_n$，它们的并是整个样本空间，可以利用下面的贝叶斯定理计算任何后验概率 $P(A_i \mid B)$。

贝叶斯定理

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n)} \tag{4-19}$$

只要知道了先验概率 $P(A_1), P(A_2), \cdots, P(A_n)$ 以及条件概率 $P(B \mid A_1), P(B \mid A_2), \cdots, P(B \mid A_n)$，就可以利用式（4-19）来计算事件 $A_1, A_2, \cdots, A_n$ 的后验概率。

4.5.1 表格法

表格法有助于完成贝叶斯定理中的计算。零件供应商问题的表格法，如表4-7所示。其中计算按以下几个步骤完成。

第一步：准备以下3列数据：

第1列——需要计算后验概率的互斥事件 A_i。

第2列——事件的先验概率 $P(A_i)$。

第3列——新信息 B 关于每个事件 A_i 的条件概率 $P(B \mid A_i)$。

第二步：在第4列中用乘法公式计算每一事件 A_i 和新信息 B 的联合概率 $P(A_i \cap B)$。联合概率由第2列的先验概率乘以第3列的条件概率得到，即 $P(A_i \cap B) = P(A_i)P(B \mid A_i)$。

第三步：把第4列中所有的联合概率加总，其和即为新信息 B 的概率 $P(B)$。在表4-7中我们看到，某零件不

⊖ 至今我们仍在使用的贝叶斯定理，是基督教长老会牧师贝叶斯（Reverend Thomas Bayes，1702～1761年）的原创性的工作。

⊜ 如果事件的并构成了整个样本空间，则称事件为完全穷尽的。

合格且来自供应商1的概率为0.0130，某零件不合格且来自供应商2的概率为0.0175。由于不合格零件只能以这两种方式获得，其和0.0130+0.0175表明在这两个供应商运来的零件中出现不合格零件的概率为0.0305。

第四步：在第5列，利用条件概率的基本关系计算后验概率。

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)}$$

注意联合概率$P(A_i \cap B)$在第4列中，而概率$P(B)$是第4列的列和。

表4-7 在两个供应商问题中计算贝叶斯定理的表格法

(1)	(2)	(3)	(4)	(5)
事件 A_i	先验概率 $P(A_i)$	条件概率 $P(B \mid A_i)$	联合概率 $P(A_i \cap B)$	后验概率 $P(A_i \mid B)$
A_1	0.65	0.02	0.0130	0.0130/0.0305 = 0.4262
A_2	0.35	0.05	0.0175	0.0175/0.0305 = 0.5738
	1.00		$P(B)$ = 0.0305	1.0000

注释和评论

1. 贝叶斯定理广泛应用于决策分析中。先验概率经常是由决策者主观估计的。在取得样本信息后，计算后验概率以供决策者选择最佳策略。
2. 任何事件和它的补都是互斥的，它们的并是整个样本空间。因此，在计算某事件和它的补的后验概率时，常常采用贝叶斯定理。

练习

方法

40. 事件A_1、A_2和A_3的先验概率分别为$P(A_1)=0.20$，$P(A_2)=0.50$，$P(A_3)=0.30$，给定A_1、A_2或A_3时事件B的条件概率为$P(B \mid A_1)=0.50$，$P(B \mid A_2)=0.40$和$P(B \mid A_3)=0.30$。
    a. 计算$P(B \cap A_1)$、$P(B \cap A_2)$和$P(B \cap A_3)$。
    b. 利用贝叶斯定理式(4-19)计算后验概率$P(A_2 \mid B)$。
    c. 采用表格法根据贝叶斯定理计算$P(A_1 \mid B)$，$P(A_2 \mid B)$和$P(A_3 \mid B)$。

应用

☆42. 某地方银行对其信用卡制度进行审核并考虑收回部分信用卡。过去，大约有5%的信用卡持有者不履行债务并造成银行无法收回的坏账。因此，管理层认为某特定的信用卡持有者不履行债务的先验概率为0.05。该银行还发现，最终履行债务的信用卡持有者以0.20的概率会拖欠到一个月后支付。当然，对于不履行债务的信用卡持有者，拖欠一个月仍未支付的概率为1。
    a. 如果某个信用卡持有者已经欠款一个月以上，计算他将不履行债务的后验概率。
    b. 如果某个信用卡持有者不履行债务的概率超过0.20，银行就将收回他的信用卡。如果某信用卡持有者已经欠款一个月，则银行是否会收回他的信用卡？为什么？

44. 据Open Doors 2011年报道，在2009～2010学年美国全日制毕业生中有9.5%的人赴海外求学（《国际教育研究》，2011年11月14日）。假设毕业记录显示，2009～2010学年赴海外求学的学生中有60%是女性，没有赴海外求学的学生中仅有49%是女性。
    a. 令A_1={学生在2009～2010学年赴海外求学}
    A_2={学生在2009～2010学年没有赴海外求学}
    W={学生为女性}
    利用给出的信息，求$P(A_1)$，$P(A_2)$，$P(W \mid A_1)$和$P(W \mid A_2)$的值。
    b. 2009～2010学年，一名女生赴海外求学的概率是多少？
    c. 2009～2010学年，一名男生赴海外求学的概率是多少？
    d. 利用前面的结果，求2009～2010学年进行全日制学习的男生和女生的百分比分别是多少？

小 结

本章，我们介绍了概率的基本概念，并举例说明如何利用概率为决策提供有用的信息。我们描述了概率如何度量事件发生可能性的大小。此外，我们看到在计算事件的概率时，既可以将构成事件的试验结果（样本点）的概率全部相加，也可以使用概率的加法公式、条件概率、乘法公式等公式。我们还演示了在获得新的信息的情况下，如何利用贝叶斯定理来得到修正的后验概率。

关键术语

probability 概率 衡量某个事件发生可能性的数值度量。

experiment 试验 一个能够产生明确结果的过程。

sample space 样本空间 由所有试验结果组成的集合。

sample point 样本点 样本空间的元素，一个样本点代表一种试验结果。

multiple-step experiment 多步骤试验 可以分为循序的多个步骤的试验。如果一个多步骤试验由 k 个步骤组成，并且在第1个步骤中有 n_1 种试验结果，在第2个步骤中有 n_2 种试验结果，……，依次类推，那么试验结果的总数为 $n_1 \times n_2 \times \cdots \times n_k$。

tree diagram 树形图 一种有助于直观描述多步骤试验的图形。

combination 组合 在从 N 个个体中选取 n 个的试验中，若不考虑个体被选取的顺序，我们对有多少种不同的选取方法感兴趣。选取的 n 个对象称为一个组合，从 N 个个体中选取 n 个个体的组合数为 $C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!}$，其中 $n=0, 1, 2, \cdots, N$。

permutation 排列 在从 N 个个体中选取 n 个的试验中，若认为个体被选取的顺序非常重要，我们对有多少种不同的选取方法感兴趣。n 个对象的选取方式称为一个排列，从 N 个个体中选取 n 个个体的排列数为 $P_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!}$，其中 $n=0, 1, 2, \cdots, N$。

basic requirements for assigning probabilities 概率分配的基本条件 概率分配必须满足如下两个基本条件：
(1) 对于任意一个试验结果 E_i，必须有 $0 \leqslant P(E_i) \leqslant 1$。
(2) 对所有 n 个试验结果，必须有 $P(E_1)+P(E_2)+\cdots+P(E_n)=1.0$。

classical method 古典法 一种概率分配的方法，适用于试验结果都是等可能发生的情形。

relative frequency method 相对频数法 一种概率分配的方法，适用于试验可以大量重复并且能取得各种试验结果发生频率的情形。

subjective method 主观法 一种以主观判断为基础的概率分配方法。

event 事件 样本点组成的集合。

complement of A 事件 A 的补 由所有不包含在 A 中的样本点构成的事件。

venn diagram 文氏图 一种抽象表明样本空间和事件运算的图形表示法，其中用矩形表示样本空间，用圆形表示事件。

union of A and B 事件 A 和事件 B 的并 由所有属于 A 或 B 或同时属于二者的样本点构成的事件，记作 $A \cup B$。

intersection of A and B 事件 A 和事件 B 的交 由同时属于 A 和 B 的样本点构成的事件，记作 $A \cap B$。

addition law 加法公式 加法公式用于计算两个事件的并的概率，有 $P(A \cup B)=P(A)+P(B)-P(A \cap B)$。对于互斥事件，由于 $P(A \cap B)=0$，此时加法公式简化为 $P(A \cup B)=P(A)+P(B)$。

mutually exclusive events 互斥事件 没有共同样本点的事件，即 $A \cap B$ 为空集，从而 $P(A \cap B)=0$。

conditional probability 条件概率 一个给定事件已经发生的条件下，另一个事件发生的概率。给定事件 B 下 A 的条件概率为 $P(A \mid B)=P(A \cap B)/P(B)$。

joint probability 联合概率 两个事件都发生的概率，即两个事件交的的概率。

marginal probability 边际概率 在联合概率表的边缘的值，表示每个事件各自发生的概率。

independent events 独立事件 若两个事件 A 和 B 独立，则 $P(A \mid B)=P(A)$ 或 $P(B \mid A)=P(B)$；也就是说，事件之间相互没有影响。

multiplication law 乘法公式 用于计算两事件交的概率公式。可以写成 $P(A \cap B)=P(A)P(B \mid A)$ 或者 $P(A \cap B)=P(B)P(A \mid B)$。对于独立事件，乘法公

式简化为 $P(A\cap B)=P(A)P(B)$。

prior probability 先验概率 事件的初始估计概率。

posterior probability 后验概率 根据新的附加信息对事件概率的修正。

bayes's theorem 贝叶斯定理 一种用来计算后验概率的方法。

重要公式

组合计数法则

$$C_n^N=\binom{N}{n}=\frac{N!}{n!(N-n)!} \quad (4\text{-}1)$$

排列计数法则

$$P_n^N=n!\binom{N}{n}=\frac{N!}{(N-n)!} \quad (4\text{-}2)$$

利用事件的补计算概率

$$P(A)=1-P(A^c) \quad (4\text{-}5)$$

加法公式

$$P(A\cup B)=P(A)+P(B)-P(A\cap B) \quad (4\text{-}6)$$

条件概率

$$P(A\mid B)=\frac{P(A\cap B)}{P(B)} \quad (4\text{-}7)$$

$$P(B\mid A)=\frac{P(A\cap B)}{P(A)} \quad (4\text{-}8)$$

乘法公式

$$P(A\cap B)=P(B)P(A\mid B) \quad (4\text{-}11)$$

$$P(A\cap B)=P(A)P(B\mid A) \quad (4\text{-}12)$$

独立事件的乘法公式

$$P(A\cap B)=P(B)P(A) \quad (4\text{-}13)$$

贝叶斯定理

$$P(A_i\mid B)=\frac{P(A_i)P(B\mid A_i)}{P(A_1)P(B\mid A_1)+P(A_2)P(B\mid A_2)+\cdots+P(A_n)P(B\mid A_n)} \quad (4\text{-}19)$$

补充练习

46. 《华尔街日报》/哈里斯个人理财民意调查中，询问2 082名成年人是否拥有自己的住房（All Business网址，2008年1月23日）。接受调查的受访者中总共有1 249人回答“是”。在450名18~34年龄组的受访者中，有117人回答“是”。
 a. 调查中的受访者拥有自己的住房的概率为多少？
 b. 在18~34年龄组的受访者中，拥有自己的住房的概率为多少？
 c. 调查中的受访者没有自己的住房的概率为多少？
 d. 在18~34年龄组的受访者中，没有自己的住房的概率为多少？

48. 43%的美国人通过大众媒体和其他网站发表他们对电视节目的评价（《赫芬顿邮报》，2011年11月23日）。对1 400名受访者进行调查，询问他们是否通过大众媒体和其他网站发表他们对电视节目的评价，调查结果如下。

		是否通过大众媒体和其他网站发表他们对电视节目的评价	
		是	否
性别	女	395	291
	男	323	355

 a. 受访者是女性的概率有多大？
 b. 假定某个受访者是一名女性，求她通过大众媒体和其他网站评价电视节目的条件概率。
 c. 令事件 F={受访者是女性}，事件 A={通过大众媒体和其他网站发表对电视节目的评价}。事件 F 和事件 A 独立吗？

50. 通过电话调查了解观众对某一新播出电视节目的反响，获得如下数据。

评价	频数	评价	频数
很差	4	高于平均水平	14
低于平均水平	8	很好	13
平均水平	11		

 a. 随机选择一名观众，他对新节目的评价为“平均水平或更好”的概率为多少？
 b. 随机选择一名观众，他对新节目的评价为“低于平均水平或更差”的概率为多少？

52. 对新注册的2 018名MBA学生进行一项调查，得到如下数据。

		申请的学校是否多于1所	
		是	否
年龄组	23岁或以下	207	201
	24~26岁	299	379
	27~30岁	185	268
	31~35岁	66	193
	36岁或以上	51	169

a. 随机选择一名MBA学生，观察学生的年龄和是否申请一所以上学校。请编制联合概率分布表。

b. 随机选择一名申请人，其年龄为23岁或23岁以下的概率为多少？

c. 随机选择一名申请人，其年龄为26岁或26岁以上的概率为多少？

d. 随机选择一名申请人，其申请学校的数目多于1所的概率为多少？

54. 搜索引擎和其他一些网站搜集个人信息并利用这些信息修正搜索结果或者发放定向广告。2012年2月，皮尤互联网与美国人生活项目（Pew Internet & American Life）进行了一次调查，调查中的一些问题涉及网络使用者对上述行为的看法（皮尤研究中心，2012年3月9日）。其中要求受访者回答的一个问题是："如果搜索引擎持续跟踪你的搜索，并根据这些信息为你未来的搜索提供个性化服务，你对此做何感想？"有两个选项供受访者选择，一种是"不赞同，感觉这样做侵犯了自己的个人隐私"，另一种是"赞同，虽然个人信息被搜集"。按年龄段分组汇总后的频率数据见下表。

年龄	不赞同	赞同
18~29	0.1485	0.0604
30~49	0.2273	0.0907
50以上	0.4008	0.0723

a. 调查中一名受访者对这种做法不赞同的概率有多大？

b. 若一名受访者的年龄在18~29岁之间，则该受访者赞同这种做法的概率有多大？

c. 若一名受访者对这种做法不赞同，则该受访者年龄在50岁以上的概率为多少？

d. 受访者对这种做法的态度是否与受访者的年龄独立？为什么？

e. 年龄在18~29岁之间的受访者和50岁以上的受访者，他们对这种做法的态度有差异吗？

56. Cooper Realty公司是位于纽约州奥尔巴尼市的一家小型房地产开发公司，主要销售民用住宅。最近，这家公司对在特定的一段时间内一座居民住宅被售出的可能性尤为关注。通过对历年该公司已售出的800套住宅的分析，得到如下数据。

		销售时间			
		30天以下	31~90天	90天以上	合计
初始定价	150 000美元以下	50	40	10	100
	150 000~199 999美元	20	150	80	250
	200 000~250 000美元	20	280	100	400
	250 000美元以上	10	30	10	50
	合计	100	500	200	800

a. 令事件A={住宅售出时间超过90天}，估计事件A的概率。

b. 令事件B={住宅定价低于150 000美元}，估计事件B的概率。

c. 事件$A \cap B$的概率为多少？

d. 假设刚刚签订的合同中某住宅的定价低于150 000美元，则Cooper Realty公司需要超过90天的时间才能将其售出的概率为多少？

e. 事件A和事件B独立吗？

58. 据一项调查报告，18岁及以上的互联网使用者中有8%的人有自己的博客。称18~29岁年龄组为青年组，调查显示：开通博客的人中54%是年轻人；没有开通博客的人中24%是年轻人（皮尤互联网与美国人生活项目，2006年7月19日）。

a. 根据这些数据，建立联合概率表。表中有2行（开通博客，未开通博客）和2列（年轻人，其他成年人）。

b. 一名互联网使用者是年轻人的概率是多少？

c. 一名互联网使用者是开通博客的年轻人的概率是多少？

d. 假定随后的一次电话调查中受访者年龄24岁，则他开通博客的概率是多少？

60. 电邮广告中，出现最为频繁的五个词分别是：shipping!，today!，here!，available!和fingertips!（《电邮广告中出现最为频繁的单词》，安迪格·林伯格，2010年3月17日）。许多电邮广告过滤器利用贝叶斯定理区分电邮广告和电邮非广告。假定对某个邮箱账户，每10封邮件中就有1封是电邮广告，每封电邮广告中上述五个最常见单词出现的频率如下：

shipping!	0.051
today!	0.045
here!	0.034
available!	0.014
fingertips!	0.014

并且假定在电邮非广告中，这五个单词出现的频率为：

shipping!	0.0015
today!	0.0022
here!	0.0022
available!	0.0041
fingertips!	0.0011

a. 若一封邮件中包括单词“shipping!”，则这是一封电邮广告的概率为多少？若一封邮件中包括单词“shipping!”，则这不是一封电邮广告的概率为多少？包括单词“shipping!”的邮件应该被标记为电邮广告吗？

b. 若一封邮件中包括单词“today!”，则这是一封电邮广告的概率为多少？若一封邮件中包括单词“here!”，则这是一封电邮广告的概率为多少？这两个单词哪一个更能说明邮件可能是一封电邮广告？为什么？

c. 若一封邮件中包括单词“available!”，则这是一封电邮广告的概率为多少？若一封邮件中包括单词“fingertips!”，则这是一封电邮广告的概率为多少？这两个单词哪一个更能说明邮件可能是一封电邮广告？为什么？

d. 从（b）和（c）中结果可见，当电邮广告滤波器利用贝叶斯定理过滤电邮广告时，什么是影响过滤效果的关键因素？

案例 Hamilton 县的法官们

Hamilton 县的法官们每年审判成千上万的案件。被处理的案件中绝大多数不会再改动其判决，但是也有一些上诉的案件，并且这些上诉的案件中，有一些判决确实被推翻。*The Cincinnati Enquirer* 的 Kristen DelGuzzi 对 Hamilton 县法官在 3 年中处理的案件进行了调查，表 4-8 中给出了由中级诉讼庭、民事庭、市政庭的 38 名法官处理过的 182 908 件案件的结果。其中有两名法官 Dinkelacker 和 Hogan 在三年之中从未在同一法庭共事。

报纸进行该研究的目的在于评估法官们的业绩。上诉通常主要是由法官的错误引起的，而该研究想了解哪些法官工作表现更好，哪些法官犯过多的错误。请你帮助分析数据，利用有关概率和条件概率的知识来对法官们的业绩进行排名。你还要分析在不同法庭处理过的案件被上诉和推翻原判的可能性有多大。

表 4-8 Hamilton 县法庭处理的所有案件以及上诉和被推翻原判的情况

法官	中级诉讼庭 处理案件总数	上诉案件数目	推翻原判案件数目
Fred Cartolano	3 037	137	12
Thomas Crush	3 372	119	10
Patrick Dinkelacker	1 258	44	8
Timothy Hogan	1 954	60	7
Robert Kraft	3 138	127	7
William Mathews	2 264	91	18
William Morrissey	3 032	121	22
Norbert Nadel	2 959	131	20
Arthur Ney, Jr.	3 219	125	14
Richard Niehaus	3 353	137	16
Thomas Nurre	3 000	121	6
Jonh O'Connor	2 969	129	12
Robert Ruehlman	3 205	145	18
J. Howard sundermann	955	60	10
Ann Marie Tracey	3 141	127	13
Ralph Winkler	3 089	88	6
合计	43 945	1762	199

（续）

法官	民事庭		推翻原判案件数目
	处理案件总数	上诉案件数目	
Penelope Cunningham	2 729	7	1
Patrick Dinkelacker	6 001	19	4
Deborah Gaines	8 799	48	9
Ronald Panioto	12 970	32	3
合 计	30 499	106	17

法官	市政庭		推翻原判案件数目
	处理案件总数	上诉案件数目	
Mike Allen	6 149	43	4
Nadine Allen	7 812	34	6
Timothy Black	7 954	41	6
David Davis	7 736	43	5
Leslie Isaiah Gaines	5 282	35	13
Karla Grady	5 253	6	0
Deidra Hair	2 532	5	0
Dennis Helmick	7 900	29	5
Timothy Hogan	2 308	13	2
James Patrick Kenney	2 798	6	1
Joseph Luebbers	4 698	25	8
William Mallory	8 277	38	9
Melba Marsh	8 219	34	7
Beth Mattingly	2 971	13	1
Albert Mestemaker	4 975	28	9
Mark Painter	2 239	7	3
Jack Rosen	7 790	41	13
Mark Schweikert	5 403	33	6
David Stockdale	5 371	22	4
John A. West	2 797	4	2
合 计	108 464	500	104

管理报告

准备一份报告对法官们的业绩进行排名。报告中要包含在三个法庭中上诉和推翻原判的可能性分析。报告至少应包括以下内容：

1. 在三个不同法庭中，案件被上诉并推翻原判的概率。
2. 每名法官所处理的案件被上诉的概率。
3. 每名法官所处理的案件被推翻原判的概率。
4. 对每名法官，当其处理的案件已被上诉时，最终推翻原判的概率。
5. 对每个法庭的法官进行排名。阐述你用于作为评定依据的标准，并说明理由。

第5章

离散型概率分布

CHAPTER 5

实践中的统计

花旗银行[⊖]
纽约州，长岛县

花旗银行是花旗集团的一个分支机构，它提供全方位的金融服务，包括支票和储蓄账户、贷款和抵押、保险以及投资服务。花旗银行通过其独一无二 Citibanking 的系统提供这些服务。

花旗银行是美国第一家引进自动提款机（ATM）的金融机构。花旗银行信用卡中心（CBC）的 ATM 机使得客户能够在任何一个地方只需用动一动手指便能瞬间完成他们的银行业务。每天24小时，每周7天，从存款到投资管理超过150种不同的银行服务项目都可以轻松完成。目前，客户80%的交易是通过 ATM 来完成的。

花旗银行信用卡中心向随机到达的客户提供服务。这是一个排队等待系统，如果遇到所有的 ATM 都忙的话，那么新来的客户只能排队等候。定期研究 CBC 能力可以分析客户的等待时间，并决定是否需要增加新的 ATM 机。

花旗银行收集的数据表明，随机到达的客户人数服从泊松分布。利用泊松分布，花旗银行可以计算任意时间段内到达 CBC 的顾客数的概率，从而决定所需 ATM 机的数目。比如，令 $x=1$ 分钟内到达的顾客数。假定到达某台 CBC 的顾客数的平均值为每分钟2人，下表给出了1分钟内到达顾客数的概率。

1分钟内到达顾客数的概率分布

x	概率	x	概率
0	0.1353	3	0.1804
1	0.2707	4	0.0902
2	0.2707	≥5	0.0527

本章主要内容是离散型概率分布，比如花旗银行分析使用的泊松分布。除泊松分布之外，我们还要介绍二项分布和超几何分布，并说明如何利用这些分布得到有用的概率信息。

本章通过介绍随机变量和概率分布的概念，将对于概率的研究进一步扩展。随机变量和概率分布是关于总体数据的模型。所谓随机变量的值表示的是数据值，概率分布给出的是取数据值时所对应的概率或者一种用于计算数据各种取值的概率的准则。本章的重点是研究离散型数据的概率分布，即离散型概率分布。

我们将介绍离散型概率分布的两类表达形式。第一种是表格形式，其中第一列是随机变量的值，第二列是随机变量取这些值时相应的概率。在第4章介绍的为试验结果分配概率的方法，可用于为分布分配概率。离散型概率分布的第二类表达形式是数学函数，计算随机变量取每种值的概率。我们给出了以函数形式表达的在实践中广泛应用的三种概率分布——二项分布、泊松分布和超几何分布。

5.1 随机变量

在第4章我们定义了试验的概念及其与之相联系的试验结果。随机变量提供了用数值描述试验结果的方法。[⊜]

> **随机变量**
> **随机变量**（random variable）是对一个试验结果的数值描述。

实际上，随机变量将每一个可能出现的试验结果赋予一个数值，随机变量的值取决于试验结果。随机变量根据取值可分为**离散型**或**连续型**。

⊖ 作者感谢花旗银行的 Stacey Karter 女士，她为实践中的统计提供了本案例。
⊜ 随机变量的取值必须是数值。

5.1.1 离散型随机变量

可以取有限多个值或无限可数多个值（如0，1，2，…）的随机变量称为**离散型随机变量**（discrete random variable）。例如，一名会计参加注册会计师（CPA）考试，考试共有4门课程。令试验中随机变量 x = 通过 CPA 考试的课程数，x 的取值可能是0，1，2，3或4，是有限个，因此 x 是一个离散型随机变量。

下面，给出离散型随机变量的另一个例子——考虑汽车到达某个收费站的试验。此时，感兴趣的随机变量 x = 一天中到达的汽车数。x 可能取整数序列 {0，1，2，…} 当中任何一个值。因此，x 是一个离散型随机变量，有无限多种取值。

尽管很多试验的结果都可以自然而然地用数值来表示，但有些试验的结果却不能。例如，在一项调查中设计的问题是“电视观众能否回忆起最近看到的一则电视广告中的信息”，试验包括两种可能的试验结果：观众能回忆起和观众不能回忆起信息。定义离散型随机变量将试验结果数值化：如果观众不能回忆起信息，则令 $x=0$；如果观众能回忆起信息，则令 $x=1$。随机变量的数据值其实是任意的（我们也可以取为5或10），但是按照随机变量的定义这些取值是可行的，即：x 给出的是对试验每个结果的数值描述，从而 x 是随机变量。

表5-1给出了其他一些离散型随机变量的例子。我们注意到：在每个例子中，离散型随机变量取有限多个或无限可数多个值（如0，1，2，…）。在本章中，我们将详细讨论这些离散型随机变量。

表5-1 离散型随机变量的例子

试验	随机变量（x）	随机变量的可能值
与5位客户洽谈	下订单的客户数	0，1，2，3，4，5
检验一批50台收音机组成的货物	损坏的收音机数	0，1，2，…，49，50
一家餐馆营业一天	顾客数	0，1，2，3…
销售一辆汽车	顾客性别	男性为0，女性为1

5.1.2 连续型随机变量

可以取某一区间或多个区间内任意值的随机变量称为**连续型随机变量**（continuous random variable）。度量时间、重量、距离、温度时，其试验结果可以用连续型随机变量来描述。例如，监控打进一家大型保险公司投诉办公室的电话的情况。假定试验中感兴趣的随机变量为 x = 相邻两个电话的间隔时间（单位：分钟）。随机变量的值可以是区间 $x \geqslant 0$ 中的任意值。事实上，x 可能取无穷多种值，比如：1.26分钟、2.751分钟、4.333分钟等。再比如，佐治亚州亚特兰大以北有一条90英里长的州际高速公路I-75。亚特兰大有一个紧急救护机构，我们可以定义随机变量 x = 在I-75公路上发生下一起事故的位置。这时，x 是一个连续型随机变量，可以取区间 $0 \leqslant x \leqslant 90$ 上的任意值。表5-2还列出了其他一些连续型随机变量的例子。注意，每一个例子中，随机变量假定可以取某个区间中的任意值。我们将在第6章介绍连续型随机变量及其概率分布。

表5-2 连续型随机变量的例子

试验	随机变量（x）	随机变量的可能值
银行开业	两客户到达的时间间隔（单位：分钟）	$x \geqslant 0$
填充一个软饮料罐（最大容量为12.1盎司）	盎司数	$0 \leqslant x \leqslant 12.1$
新建一座图书馆	六个月后工程进度	$0 \leqslant x \leqslant 100$
检验一种新的化学方法	反应发生的温度（最低150 ℉，最高212 ℉）	$150 \leqslant x \leqslant 212$

注释和评论

一种确定随机变量是离散型还是连续型的方法，是把随机变量的值看作一条线段上的点。任意选择随机变量

的两个值，假如线段上这两点之间的所有点都可能是随机变量的取值，则该随机变量就是连续型的。

练习

方法

2. 考虑工人组装产品的试验。
   a. 定义一个随机变量，用以表示组装产品所需的的时间（单位：分钟）。
   b. 随机变量可能取哪些值?
   c. 这个随机变量是离散型还是连续型?

应用

4. 1月美国失业率下降到为8.3%（美国劳工部网站，2012年2月10日）。美国人口普查局从东北部地区选取九个州。假定我感兴趣的随机变量是，东北部地区九个州中1月的失业率低于8.3%的州的数目。试问，这个随机变量可能取哪些值?

6. 以下是一系列试验及相关的随机变量。在每一个试验中，确定随机变量的取值，并说明随机变量是离散型随机变量还是连续型随机变量。

试验	随机变量（x）
a. 参加一个20道题的考试	回答正确的问题数
b. 观察1小时中到达收费站的汽车	到达收费站的汽车数
c. 审计50份税务报告	出现错误的报告数
d. 观察一名雇员的工作	在8小时工作日中非生产性工作的小时数
e. 称一批货物的重量	磅数

5.2 离散型概率分布

随机变量的**概率分布**（probability distribution）是描述随机变量取不同值的概率。对于离散型随机变量x，**概率函数**（probability function）给出随机变量取每种值的概率，记做$f(x)$。在建立离散型概率分布时，可以采用第四章中所介绍的分配概率的方法——古典法、主观法和相对频率法。本节中，我们演示如何使用这些方法。用这些方法可以得到所谓的离散型概率分布表，即将概率分布用表格形式给出。

当各种试验结果对应的随机变量值是等概率时，适合采用古典法为随机变量的值分配概率。比如，在抛掷一枚色子的试验中观测正面朝上一面的点数。可能出现的点数为1，2，3，4，5或6，并且每种结果发生的可能性是均等的。因此，若令x = 抛掷一枚色子正面朝上的点数，$f(x)=x$的概率，则x的概率分布见表5-3。

采用主观法为试验结果分配概率也会得到一张表，其中有随机变量的取值及其相应概率。主观法中，每个人根据自己的最优判断分配概率，建立概率分布。因此，与根据法建立概率分布不同，主观法得到的概率分布可能因人而异，不同的人会给出不同的概率分布。

表5-3 抛掷一枚色子朝上一面点数的概率分布

x	$f(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

当数据量相当大的时候，可以采用相对频率法为随机变量的值分配概率。这时，我们将数据看作总体，采用相对频率法为试验结果分配概率。采用相对频率法建立离散型概率分布得到所谓的**经验离散分布**（empirical discrete distribution）。如今很容易获取大量的数据（比如：扫描数据，信用卡数据），这使得这类分布在实际中广为应用。下面，我们以汽车代理商的销售量为例举例说明其应用。

采用相对频率法，我们建立DiCarlo公司在纽约萨拉托加的汽车销售量的概率分布。历史数据显示，在过去300天的营业时间中，有54天汽车销售量为0辆，117天为1辆，72天为2辆，42天为3辆，12天为4辆，3天为5辆。假设我们观测DiCarlo公司一天的经营情况，试验中定义随机变量x = 汽车的日销售量。采用相对频率法为随机变量x的值分配概率，建立x的概率分布。

在概率函数的表达式中，$f(0)$表示销售0辆汽车的概率，$f(1)$表示销售1辆汽车的概率，以此类推。由于

历史数据显示，300 天中有 54 天汽车销售量为 0，于是将相对频率 54/300 = 0.18 分配给 $f(0)$，表示一天中销售 0 辆汽车的概率为 0.18。同理，由于历史数据显示 300 天中有 117 天销售 1 辆汽车，于是将相对频率 117/300 = 0.39 分配给 $f(1)$，表示一天中销售 1 辆汽车的概率为 0.39。对随机变量的其他值，继续采用这种方法，计算得到 $f(2)$，$f(3)$，$f(4)$，$f(5)$ 的值，见表 5-4。

表 5-4 DiCarlo 公司汽车日销售量的概率分布

x	$f(x)$
0	0.18
1	0.39
2	0.24
3	0.14
4	0.04
5	0.01
	合计 1.00

定义一个随机变量及其概率分布的最基本的有利条件在于，一旦掌握概率分布则确定决策者感兴趣的各种事件的概率就变得相当容易了。例如，根据表 5-4 中 DiCarlo 汽车公司的概率分布可知，一天当中最有可能卖出的汽车数量为 1，其概率 $f(1) = 0.39$。另外，汽车日销售量大于或等于 3 的概率是 $f(3) + f(4) + f(5) = 0.14 + 0.04 + 0.01 = 0.19$。这些概率及其他信息有助于决策者掌握 DiCarlo 公司的汽车销售情况。

一个离散型随机变量的概率函数必须满足如下两个条件[⊖]：

离散型概率函数的基本条件

$$f(x) \geq 0 \tag{5-1}$$

$$\sum f(x) = 1 \tag{5-2}$$

表 5-4 表明，随机变量 x 的概率满足式（5-1），即：对于任意 x，$f(x)$ 大于或等于零。另外，概率和为 1，故满足式（5-2）。于是，DiCarlo 公司的概率函数是一个有效的离散型概率函数。

我们还可以用图形来表示 DiCarlo 汽车公司的概率分布。在图 5-1 中，横轴坐标表示 DiCarlo 公司随机变量 x 的值，纵轴坐标取这些值相应的概率。

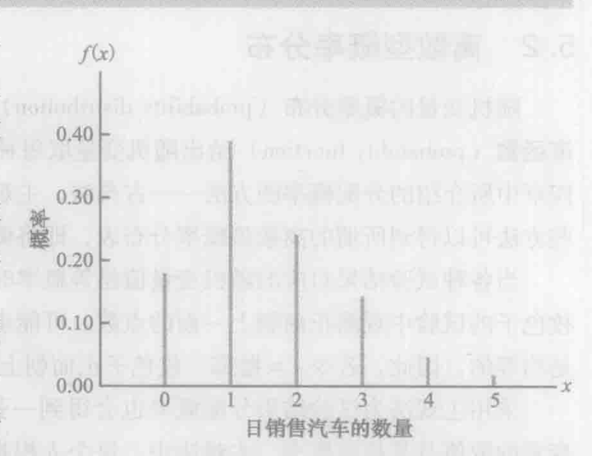


图 5-1 DiCarlo 公司汽车日销售量的概率分布图

除了表中的概率分布，还可以用公式给出随机变量 x 的概率函数 $f(x)$。其中，$f(x)$ 给出随机变量 x 取每种值的概率。在离散型概率分布中，**离散型均匀概率分布**（discrete uniform probability distribution）是最简单的一类，其概率函数的定义见下式（5-3）。

离散型均匀概率函数

$$f(x) = 1/n \tag{5-3}$$

式中，n 代表随机变量可能取值的个数。

例如，考虑抛一枚色子的试验，定义随机变量 x 为色子面向上的点数。对于这个试验，随机变量有 $n = 6$ 个可能值，即 $x = 1, 2, 3, 4, 5$ 和 6。由于每个值出现的概率是相等的，因此离散型均匀随机变量 x 的概率函数为：

$$f(x) = 1/6 \qquad x = 1,2,3,4,5,6$$

常用离散型随机变量的概率分布通常是以公式的形式给出，二项分布、泊松分布和超几何分布是其中最重要的三类分布，本章随后将介绍这三类分布。

⊖ 这些条件与第 4 章中对试验结果分配概率时提出的两条基本要求类似。

练 习

应用

☆8. 以下数据是坦帕总医院20天中手术室的使用情况：有3天只使用1间，有5天使用2间，有8天使用3间，有4天医院的4间手术室都被使用。

a. 根据相对频数法，对一天中手术室的使用情况建立经验离散概率分布。

b. 绘制概率分布图。

c. 说明这个概率分布满足离散型概率分布的条件。

10. 从信息系统（IS）选取高级经理和中级经理组成一个样本，下表是他们工作满意度得分的百分数频数分布表。其中，工作满意度得分从1（非常不满意）到5（非常满意）。

a. 编制高级经理工作满意度得分的概率分布。

b. 编制中级经理工作满意度得分的概率分布。

c. 一名高级经理的工作满意度为4或5的概率是多少？

d. 中级经理对工作非常满意的概率是多少？

e. 对比高级经理和中级经理整体工作满意度。

工作满意度得分	高级经理（%）	中级经理（%）
1	5	4
2	9	10
3	3	12
4	42	46
5	41	28

12. 康卡斯特有线传播和时代华纳有限公司是两家最大的有线电视供应商，他们分别有2 150万用户和1 100万用户（《纽约时报年鉴》，2007年）。假定时代华纳有限传播的管理人员主观估计明年纽约州新用户数的概率分布如下表所示。

x	$f(x)$	x	$f(x)$
100 000	0.10	400 000	0.30
200 000	0.20	500 000	0.10
300 000	0.25	600 000	0.05

a. 这是一个概率分布吗？请说明理由。

b. 时代华纳有线新用户超过400 000户的概率是多少？

c. 时代华纳有线新用户少于200 000户的概率是多少？

14. 下表是MRA公司第一年营业时预计利润x（单位：1 000美元）的概率分布，其中，负值代表亏损。

x	$f(x)$	x	$f(x)$
−100	0.10	100	0.25
0	0.20	150	0.10
50	0.30	200	

a. $f(200)$ 的值是多少？你怎样解释这个值？

b. MRA赢利的概率是多少？

c. MRA至少赢利100 000美元的概率是多少？

5.3 数学期望与方差

5.3.1 数学期望

随机变量的数学期望（expected value）或均值是对随机变量中心位置的一种度量。离散型随机变量x的数学期望的数学表达式如式（5-4）所示。

离散型随机变量的数学期望[⊖]

$$E(x) = \mu = \sum x f(x) \tag{5-4}$$

通常采用记号$E(x)$或μ表示随机变量的数学期望。

式（5-4）表明，计算离散型随机变量的数学期望时，首先将随机变量的所有可能取值乘以相应的概率值$f(x)$，然后再将结果相加。利用第5.2节中DiCarlo公司汽车销售量的例子，我们演示如何计算汽车日销售量

⊖ 数学期望是随机变量取值的加权平均，其中的权数是概率。

的数学期望，见表5-5。表5-5 中 $xf(x)$ 列的总和等于1.5，即汽车日销售量的数学期望为1.5 辆。[㊀]虽然一天当中可能售出0，1，2，3，4或5辆汽车，但是在一段时间里，可以预期 DiCarlo 公司平均日销售 1.50 辆汽车。假设一个月营业30天，我们根据数学期望1.5辆/日，可以预计汽车的月平均销售量为 30×1.5=45 辆。

表5-5 计算 DiCarlo 公司汽车日销售量的数学期望

x	$f(x)$	$xf(x)$
0	0.18	0×0.18=0.00
1	0.39	1×0.39=0.39
2	0.24	2×0.24=0.48
3	0.14	3×0.14=0.42
4	0.04	4×0.04=0.16
5	0.01	5×0.01=0.05
		1.50
		$E(x)=\mu=\sum xf(x)$

5.3.2 方差

虽然数学期望给出了随机变量中心位置的度量，我们常常还需要度量随机变量的变异性或分散程度。正如我们在第3章中描述数据的变异性那样，现在用**方差**（variance）来描述随机变量取值的变异性。离散型随机变量方差的数学表达式如下。

离散型随机变量的方差

$$\mathrm{Var}(x)=\sigma^2=\sum(x-\mu)^2f(x) \tag{5-5}$$

如式（5-5）所示，方差公式的关键是离差（$x-\mu$）——度量随机变量的某一特定值与数学期望或均值μ的距离。在计算随机变量的方差时，随机变量所有值的离差平方的加权求和称为方差，通常用记号 $\mathrm{Var}(x)$ 或 σ^2 来表示随机变量的方差[㊁]。

表5-6 计算 DiCarlo 公司汽车日销售量的方差

x	$x-\mu$	$(x-\mu)^2$	$f(x)$	$(x-\mu)^2f(x)$
0	0-1.50=-1.50	2.25	0.18	2.25×0.18=0.4050
1	1-1.50=-0.50	0.25	0.39	0.25×0.39=0.0975
2	2-1.50=0.50	0.25	0.24	0.25×0.24=0.0600
3	3-1.50=1.50	2.25	0.14	2.25×0.14=0.3150
4	4-1.50=2.50	6.25	0.04	6.25×0.04=0.2500
5	5-1.50=3.50	12.25	0.01	12.25×0.01=0.1225
				1.2500
				$\sigma^2=\sum(x-\mu)^2f(x)$

在表5-6中，根据DiCarlo公司汽车日销售量概率分布，总结了计算方差的过程。可见，方差为1.25。定义方差的算术平方根为**标准差**（standard deviation），记做σ。汽车日销售量的标准差为：

$$\sigma=\sqrt{1.25}=1.118$$

标准差的单位与随机变量的单位相同（σ=1.118 辆），所以更加常用于描述一个随机变量的变异性。方差σ^2的单位是随机变量的单位的平方，其含义较难解释。

㊀ 随机变量的数学期望不一定是随机变量的某个值。

㊁ 方差是随机变量离差平方的加权算术平均，其中的权数是概率。

练 习

方法

☆16. 下表是随机变量 y 的概率分布：

y	$f(y)$	y	$f(y)$
2	0.20	7	0.40
4	0.30	8	0.10

a. 计算 $E(y)$。 b. 计算 $\text{Var}(y)$ 和 σ。

应用

☆18. 美国住房调查报告称，中心城市自有住房和租赁住房中卧室数目的数据见下表是（美国人口署网站，2003年3月31日）：

卧室数	房屋数目（单位：千套）	
	租赁住房	自有住房
0	547	23
1	5 012	541
2	6 100	3 832
3	2 644	8 690
≥4	557	3 783

a. 令随机变量 x 表示租赁住房的卧室间数（$x=4$ 表示有4间及4间以上卧室）。编制随机变量 x 的概率分布。

b. 计算租赁住房中卧室间数的数学期望和方差。

c. 令随机变量 y 表示自有住房的卧室间数（$y=4$ 表示有4间及4间以上卧室）。编制随机变量 y 的概率分布。

d. 计算自有住房中卧室间数的数学期望和方差。

e. 比较自有住房和租赁住房的卧室间数，你有什么发现？

20. 据 Newton 汽车保险公司称，碰撞事故保险损害赔付的概率分布见下表：

赔付金额（美元）	概率	赔付金额（美元）	概率
0	0.85	5 000	0.02
500	0.04	8 000	0.01
1 000	0.04	10 000	0.01
3 000	0.03		

a. 根据碰撞事故赔付金额的数学期望，确定公司止损的碰撞保险费金额。

b. 对车辆碰撞险，保险公司每年收取520美元的保险费，求保单持有人获取保险收益的数学期望是多少（提示：数学期望是指从保险公司取得的期望赔付金额减去碰撞险的成本）？有这一数学期望下，为什么客户会购买这种碰撞事故保险？

22. Carolina Industries 公司的产品月需求量存在明显差异。根据过去两年的数据，得到公司月需求量的概率分布，见下表。

需求量	概率	需求量	概率
300	0.20	500	0.35
400	0.30	600	0.15

a. 如果公司根据月需求量的数学期望来确定每月的订货量，该公司这种产品的每月订货量是多少？

b. 假设每单位产品的销售收入为70美元，订货成本为50美元。如果基于（a）中的结果订购，并且该产品的实际需求量为300，那么公司一个月的赢利或亏损是多少？

24. J. R. Ryland 计算机公司正在考虑一项厂房扩建计划，以便公司能够开始生产一种新的计算机产品。公司总裁必须决定是进行中型还是大型扩建工程。新产品的需求量是一个不确定因素，对于规划目标可能出现低、中或高三种需求，估计与这三种需求对应的概率分别为0.20、0.50和0.30。在下表中，x 和 y 分别表示策划者预测中型和大型扩建工程的年利润（单位：1000美元）。

		中型扩建工程利润		大型扩建工程利润	
		x	$f(x)$	y	$f(y)$
	低	50	0.20	0	0.20
需求	中	150	0.50	100	0.50
	高	200	0.30	300	0.30

a. 计算两种扩建方案的利润的数学期望。基于期望利润最大化的目标，你推荐哪种方案？

b. 计算两种扩建方案的利润的方差。基于风险或不确定性最小化的目标，你推荐哪种方案？

5.4 二元分布、协方差和金融资产组合

关于两个随机变量的概率分布称为**二元概率分布**（bivariate probability distribution）。在讨论二元概率分布时，

考虑一个二元试验是非常有用的。在一个二元试验中，每种试验结果由两个值构成，其中每个值与一个随机变量相对应。比如，在一个抛掷一对色子的二元试验中，试验结果由两个值构成，其中一个是第一枚色子的点数，另一个是第二枚色子的点数。再比如，观察金融市场上一只股票基金和一只债券基金，记录它们在一年中的收益率。试验结果给出随机变量的一对值，其中一个值是股票基金的收益率，另一个值是债券基金的收益率。在处理二元概率分布时，我们往往感兴趣的是随机变量之间的关系。本节，我们将介绍二元概率分布并展示如何利用协方差和相关系数度量随机变量之间线性关系的强弱。我们还会介绍如何利用二元概率分布构造和分析金融资产组合。

5.4.1 二元经验离散型概率分布

在第5.2节，我们给出了DiCarlo汽车公司在纽约萨拉加图汽车日销售量的经验离散分布。DiCarlo汽车公司在日内瓦还有另一家代理商。表5-7中是日内瓦和萨拉加图两位代理商300天的日汽车销售量。表中最后一行（合计）中的数值，是我们在第5.2节用于建立DiCarlo汽车公司在萨拉加图代理商的汽车日销售量经验概率分布时所用的频数。最后一列（合计）中的数值，是日内瓦代理商的日销售量频数。表中的数据是天数，这时，日内瓦代理商的日销售量为数据位于的行与第一列对应的那个水平，萨拉加图代理商的日销售量为数据位于的列与第一行对应的水平。比如，表中的数据33，对应的日内瓦代理商所在的行标号为1，萨拉加图代理商所在的列标号为2，这说明在300天中有33天萨拉加图的日销售量为2和日内瓦代理商的日销售量为1。

表5-7 300天中DiCarlo公司在日内瓦和萨拉加图代理商的日汽车销售量

日内瓦代理商	萨拉加图代理商						合计
	0	1	2	3	4	5	
0	21	30	24	9	2	0	86
1	21	36	33	18	2	1	111
2	9	42	9	12	3	2	77
3	3	9	6	3	5	0	26
合计	54	117	72	42	12	3	300

假定我们观察并记录DiCarlo公司一天中的汽车销售情况，这构成一个二元试验。令 x = 日内瓦代理商售出的汽车数量，y = 萨拉加图代理商售出的汽车数量。将表5-7中的数据除以观测值总数（300）得到DiCarlo公司两家代理商汽车销量的二元经验离散概率分布，见表5-8。其中，最后一行给出的是DiCarlo公司萨拉加图代理商的边际分布，最后一列给出的是DiCarlo公司日内瓦代理商的边际分布。

表中间部分的概率值是两家代理商销量的二元概率分布。二元概率分布常称作联合概率。可见，一天当中，日内瓦售出0辆车萨拉加图售出1辆车的联合概率为 $f(0, 1) = 0.1000$，日内瓦售出1辆车萨拉加图售出4辆车的联合概率为 $f(1, 4) = 0.0067$，…，以此类推。注意，每种试验结果对应着一个二元概率。x 有4个可能值，y 有6个可能值，因此有24种试验结果和二元概率。

表5-8 DiCarlo公司在纽约州日内瓦和萨拉加图两家代理商汽车日销量的二元经验离散概率分布

日内瓦代理商	萨拉加图代理商						合计
	0	1	2	3	4	5	
0	0.0700	0.1000	0.0800	0.0300	0.0067	0.0000	0.2867
1	0.0700	0.1200	0.1100	0.0600	0.0067	0.0033	0.3700
2	0.0300	0.1400	0.0300	0.0400	0.0100	0.0067	0.2567
3	0.0100	0.0300	0.0200	0.0100	0.0167	0.0000	0.0867
合计	0.18	0.39	0.24	0.14	0.04	0.01	1.0000

假定我们想要知道DiCarlo公司两家代理商销售总量的概率分布、数学期望和方差。可以定义销售总量 $s = x + y$，根据表5-8中的二元概率，有：$f(s=0) = 0.0700$，$f(s=1) = 0.0700 + 0.1000 = 0.1700$，$f(s=2) = 0.0300 + 0.1200 + 0.0800 = 0.2300$，…，以此类推。表5-9给出了 $s = x + y$ 的完整的概率分布，并计算出数学期望 $E(s) = 2.6433$ 和方差 $\mathrm{Var}(s) = 2.3895$。

表 5-9 计算 DiCarlo 汽车公司日销售总量的数学期望和方差

s	$f(s)$	$sf(s)$	$s-E(s)$	$[s-E(s)]^2$	$[s-E(s)]^2f(s)$
0	0.070 0	0.000 0	−2.643 3	6.987 2	0.489 1
1	0.170 0	0.170 0	−1.643 3	2.700 5	0.459 1
2	0.230 0	0.460 0	−0.643 3	0.413 9	0.095 2
3	0.290 0	0.870 0	0.356 7	0.127 2	0.036 9
4	0.126 7	0.506 7	1.356 7	1.840 5	0.233 1
5	0.066 7	0.333 3	2.356 7	5.553 9	0.370 3
6	0.023 3	0.140 0	3.356 7	11.267 2	0.262 9
7	0.023 3	0.163 3	4.356 7	18.980 5	0.442 9
8	0.000 0	0.000 0	5.356 7	28.693 9	0.000 0
		$E(s)=2.643\ 3$			$\mathrm{Var}(s)=2.389\ 5$

根据二元概率分布，我们常常还想知道两个随机变量之间的关系。协方差和相关系数用来度量两个随机变量之间的关系。在第 3 章中，我们介绍了如何根据样本数据计算协方差和相关系数。计算两个随机变量 x 和 y 的协方差的公式如下。

随机变量 x 和 y 的协方差[⊖]

$$\sigma_{xy}=[\mathrm{Var}(x+y)-\mathrm{Var}(x)-\mathrm{Var}(y)]/2 \tag{5-6}$$

$\mathrm{Var}(s)=\mathrm{Var}(x+y)$ 已知，并且在第 5.2 节中已经计算出 $\mathrm{Var}(y)$。在利用式（5-6）计算 x 和 y 的协方差时，还需要计算 $\mathrm{Var}(x)$。根据 x 的概率分布（见表 5-8 右侧），在表 5-10 中计算数学期望 $E(x)$ 和方差 $\mathrm{Var}(x)$。

表 5-10 计算 DiCarlo 公司日内瓦汽车代理商日销量的数学期望和方差

x	$f(x)$	$xf(x)$	$x-E(x)$	$[x-E(x)]^2$	$[x-E(x)]^2f(x)$
0	0.286 7	0.000 0	−1.143 5	1.307 6	0.374 9
1	0.370 0	0.370 0	−0.143 5	0.020 6	0.007 6
2	0.256 7	0.513 4	0.856 5	0.856 5	0.188 3
3	0.086 7	0.260 1	1.856 5	1.856 5	0.298 8
		$E(x)=1.143\ 5$			$\mathrm{Var}(x)=0.869\ 6$

利用式（5-6）计算随机变量 x 和 y 的协方差，得

$$\sigma_{xy}=[\mathrm{Var}(x+y)-\mathrm{Var}(x)-\mathrm{Var}(y)]/2=(2.389\ 5-0.869\ 6-1.25)/2=0.135\ 0$$

协方差等于 0.135 0 表明 DiCarlo 公司两家代理商的日销量正相关。要更精确地度量二者相关关系的强弱，还可以计算相关系数。

两个随机变量 x 和 y 的相关系数的计算公式见式（5-7）。由式（5-7）可见，两个随机变量的相关系数等于协方差除以两个随机变量的标准差之积。

随机变量 x 和 y 的相关系数

$$\rho_{xy}=\frac{\sigma_{xy}}{\sigma_x\sigma_y} \tag{5-7}$$

下面，计算 DiCarlo 公司两家代理商日销量的相关系数。首先，将萨拉加图和日内瓦代理商销售量的方差开算术平方根求标准差。

$$\sigma_x=\sqrt{0.869\ 6}=0.932\ 5$$

$$\sigma_y=\sqrt{1.25}=1.118\ 0$$

⊖ 当 $\mathrm{Var}(x+y)$ 未知时，另一个计算 x 和 y 协方差的常用公式为：$\sigma_{xy}=\sum_{i,j}[x_i-E(x_i)][y_j-E(y_j)]f[(x_i,y_j)]$

然后计算度量两个随机变量线性关系强弱的相关系数。

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{0.1350}{0.9325 \times 1.1180} = 0.1295$$

在第3章中，我们已将给出了相关系数的定义。相关系数度量两个变量之间线性关系的强弱。相关系数的值越接近于+1，表明同向线性关系越强；越接近于-1，表明反向线性关系越强；越接近于0，越表明没有线性关系。对于随机变量，这个性质仍然成立。相关系数为0.1295，表明两个随机变量之间有弱的正相关关系，其中随机变量是两家DiCarlo公司代理商的日销量。若相关系数为0，则我们可以得出两家代理商的日销量之间是相互独立的结论。

5.4.2 金融上的应用

下面，我们看看如何将所学的内容用于建立一个金融资产组合，在收益和风险之间寻求平衡。一名理财顾问认为来年的经济形势可能有四种情形。令 x 表示大型股票基金的投资收益率，y 表示政府长期债券基金的投资收益率。针对每种经济形势，理财顾问建立了 x 和 y 的概率分布。表5-11是 x 和 y 的二元概率分布。表中为每种试验结果（经济形势）单独列出一行。每行中包括试验结果的联合分布以及随机变量的取值。由于只有四个联合概率，因此表5-11的形式比DiCarlo汽车公司采用的表格形式要简单，DiCarlo汽车公司采用的表格中有4×6=24个联合概率。

表5-11 大型股票基金的投资收益率 x 和政府长期债券基金的投资收益率 y 的概率分布

经济形势	概率 $f(x, y)$	大型股票基金 x	长期政府债券基金 y
衰退	0.10	-40	30
缓慢增长	0.25	5	5
稳定增长	0.50	15	4
快速增长	0.15	30	2

利用第5.3节中计算单个随机变量的数学期望的公式，可以计算股票基金投资的期望收益率 $E(x)$ 和债券基金投资的期望收益率 $E(y)$。

$$E(x) = 0.1 \times (-40) + 0.25 \times 5 + 0.5 \times 15 + 0.15 \times 30 = 9.25$$

$$E(y) = 0.1 \times 30 + 0.25 \times 5 + 0.5 \times 4 + 0.15 \times 2 = 6.55$$

根据这一信息，我们可能会得出结论认为投资于股票更好。其收益率更高，达到9.25%。但是理财分析师建议投资者还要考虑投资的风险。收益率的标准差常常用于衡量风险。要计算标准差，必须先计算方差。利用第5.3节中计算单个随机变量的方差的公式，可以计算股票基金投资收益率和债券基金投资收益率的方差。

$$\begin{aligned}\text{Var}(x) &= 0.1 \times (-40 - 9.25)^2 + 0.25 \times (5 - 9.25)^2 + 0.5 \times (15 - 9.25)^2 + 0.15 \times (30 - 9.25)^2 \\ &= 328.1875\end{aligned}$$

$$\begin{aligned}\text{Var}(y) &= 0.1 \times (30 - 6.55)^2 + 0.25 \times (5 - 6.55)^2 + 0.5 \times (4 - 6.55)^2 + 0.15 \times (2 - 6.55)^2 \\ &= 61.9475\end{aligned}$$

股票基金投资收益率的标准差为 $\sigma_x = \sqrt{328.1875} = 18.1159\%$，债券基金投资收益率的标准差为 $\sigma_y = \sqrt{61.9475} = 7.8707\%$。于是，我们得出结论认为债券投资的风险更低，其标准差更小。可见，股票基金的投资收益率更高。是投资于债券基金还是投资于债券基金？这取决于你面对收益和风险的态度。一个积极的投资者可能会选择股票基金，因为它的收益率更高。一个保守的投资者可能会选择债券基金，因为它的风险更高。但是，除此之外还有其他的选择。可以选择投资于一个由股票基金和债券基金组成的资产组合。

假定我们想要考虑三种投资方式：单独投资于大型股票基金、单独投资于政府长期债券基金、将资金平均分配对股票基金和债券基金各投资一半。我们已经计算出了单独投资于股票基金和单独投资于债券基金时收益的数学期望值和标准差。现在评价第三种投资方式，建立一个资产组合，将资金平均分配对大型股票基金和政府长期债券基金各投资一半。

为了评价这个资产组合，我们从计算它的期望收益开始。前面我们已经定义 x 表示股票基金的投资收益率，y 表示债券基金的投资收益率。因此资产组合的投资收益率为 $r = 0.5x + 0.5y$。在一个资产组合中，一半投资于股票

基金一半投资于债券基金各投资一半，要计算其期望收益等价于计算 $E(r)=E(0.5x+0.5y)$。表达式 $0.5x+0.5y$ 称作随机变量 x 和 y 的线性组合。当已知 $E(x)$ 和 $E(y)$ 已知时，式（5-8）给出了一个便于计算随机变量 x 和 y 的线性组合的数学期望的方法。在式（5-8）中，a 是线性组合中 x 的系数，b 是线性组合中 y 的系数。

随机变量 x 和 y 的线性组合的数学期望

$$E(ax+by)=aE(x)+bE(y) \tag{5-8}$$

由于已经计算出 $E(x)=9.25$ 和 $E(y)=6.55$，因此可以利用式（5-8）计算资产组合的期望值。

$$E(0.5x+0.5y)=0.5E(x)+0.5E(y)=0.5\times0.25+0.5\times6.55=7.9$$

可见，资产组合的期望投资收益为7.9%。若投资100美元，则预期会收益 $100\times0.079=7.9$ 美元；若投资1 000美元，则预期会收益 $1\,000\times0.079=79.00$ 美元，…，以此类推。但是，风险如何呢？正如前文所提到的，理财分析人员常常用标准差度量风险。

我们的资产组合是两个随机变量的一个线性组合，因此为了评价资产组合的风险我们需要计算两个随机变量的线性组合的方差和标准差。当两个随机变量的协方差已知时，可以利用式（5-9）计算两个随机变量的线性组合的方差。

两个随机变量的线性组合的方差

$$\mathrm{Var}(ax+by)=a^2\mathrm{Var}(x)+b^2\mathrm{Var}(y)+2ab\sigma_{xy} \tag{5-9}$$

式中，σ_{xy} 为 x 和 y 的协方差。

由式（5-9）可见，资产组合是两个随机变量的线性组合，为了计算资产组合的方差，需要计算每个随机变量的方差以及两个随机变量之间的协方差。

我们已经计算出两个随机变量各自的方差分别为：$\mathrm{Var}(x)=328.1875$ 和 $\mathrm{Var}(y)=61.9475$。并且可以得出 $\mathrm{Var}(x+y)=119.46$ [⊖]。因此，利用式（5-6）计算两个随机变量 x 和 y 的协方差。

$$\sigma_{xy}=[\mathrm{Var}(x+y)-\mathrm{Var}(x)-\mathrm{Var}(y)]/2=(119.46-328.1875-61.9745)/2=-135.3375$$

像上文中这样得到 x 和 y 的协方差为负，意味着当 x 倾向于高于其均值时 y 倾向于低于其均值，反之亦然。

可以利用式（5-9）计算资产组合收益的方差。

$$\mathrm{Var}(0.5x+0.5y)=0.5^2\times328.1875+0.5^2\times61.9475+2\times0.5\times0.5\times(-135.3375)=29.865$$

资产组合的标准差 $\sigma_{0.5x+0.5y}=\sqrt{29.865}=5.4650\%$，它度量了股票基金和债券基金各投资50%的资产组合的风险。

对三种投资方式：单独投资于股票基金，单独投资于债券基金，建立一个资产组合将资金等额地投资于股票和债券基金，也许我们想要进行比较，表5-12给出了每种投资方式收益的数学期望、标准差和方差。

表5-12　三种投资方式的数学期望、方差和标准差

投资方式	收益率的数学期望（%）	收益率的方差	收益率的标准差（%）
100%投资于股票基金	9.25	328.187 5	18.115 9
100%投资于债券基金	6.55	61.947 5	7.870 7
资本组合（股票基金和债券基金各占50%）	7.90	29.865	5.465 0

这些方法中你更倾向于采用哪一种？100%投资于股票基金的期望收益最高，但是风险也最大，标准差为18.115 9%。100%投资于债券基金的期望收益最低，但是风险也显然更小。50%投资于股票基金50%投资于债券基金（资本组合）的期望收益是单独投资于股票基金和单独投资于债券基金的收益的平均值。但是，资产组合的风险比单独投资于任意一种基金的风险都要低。的确，资产组合比单独投资于债券基金的收益高并且风险低（标准差更小）。因此，我们说投资于资产组合优于单独投资于债券基金。

⊖ 采用上一节中在DiCarlo汽车公司例子相同的方法，计算得到 $\mathrm{Var}(x+y)=119.46$。

是选择投资于股票基金，或者债券基金还是资本组合？这取决于你面对风险的态度。股票基金的期望收益更高。不过，资本组合明显风险更低而且还有相当不错的收益。许多人愿意选择资本组合。正是由于股票基金和债券基金之间的协方差为负，使得资本组合的风险比单独投资于某一种基金的风险低这么多。

在刚才的资产组合分析中，是50%投资于股票基金50%投资于债券基金。对其他形式的资产组合，该如何计算其数学期望和方差呢？式（5-8）和式（5-9）可以方便地用于完成这些计算。

假定我们新建一个资产组合，25%投资于股票基金，75%投资于债券基金。新的资产组合的数学期望和方差是多少呢？该资产组合的收益为 $r=0.25x+0.75y$，可以根据式（5-8）计算资产组合的数学期望。

$$E(0.25x+0.75y)=0.25E(x)+0.75E(y)=0.25\times 9.25+0.75\times 6.55=7.225$$

同理，我们可以根据式（5-9）计算资产组合的方差：

$$\begin{aligned}\mathrm{Var}(0.25x+0.75y)&=0.25^2\mathrm{Var}(x)+0.75^2\mathrm{Var}(y)+2\times 0.25\times 0.75\sigma_{xy}\\&=0.0625\times 328.1875+0.5625\times 61.9475+0.375\times(-135.3375)\\&=4.6056\end{aligned}$$

新的资产组合的标准差 $\sigma_{0.25x+0.75y}=\sqrt{4.6056}=2.1461$。

5.4.3 小结

在本节，我们介绍了二元离散概率分布。由于这种分布涉及两个随机变量，我们常常对度量变量之间的关系感兴趣。我们介绍了协方差和相关系数这两种度量方法，并演示了如何计算。相关系数越接近1或者 -1，表明两个随机变量之间的关系越密切。相关系数取值接近0，则表明两个随机变量之间的关系薄弱。若两个随机变量相互独立，则协方差和相关系数都等于0。

我们还展示了如何计算随机变量的线性组合的数学期望和方差。从统计的观点来看，金融资产组合是随机变量的线性组合。实际上，这是一种我们称之为加权平均的特殊的线性组合。组合系数非负并且系数之和为1。在我们给出的例子中，资产组合由股票基金投资和债券基金投资组成。以此为例我们展示了如何计算资产组合的数学期望和方差。对于任何由两种金融资本构成的资产组合，都可以用这种方法计算其数学期望和方差。各个随机变量之间的协方差为负，这减小了资产组合的方差，大部分降低资产组合风险的理论正是基于此而提出分散投资的。

注释和评论

1. 式（5-8）和式（5-9）以及它们扩展到三个或者三个以上随机变量的结果，是金融资产组合构建和分析的关键构建。
2. 式（5-8）和式（5-9）用于计算两个随机变量的线性组合的数学期望和方差，可以扩展到三个或者三个以上随机变量的情况。式（5-8）的扩展很直接，只要在公式中加入新添加的随机变量的项目就可以了。式（5-8）的扩展则更为复杂，要在公式中加入所有随机变量的两两协方差。这些扩展留待高级教材中介绍。
3. 式（5-9）中的协方差说明了为什么负相关的随机变量（不同的投资方式）能够减小方差，从而降低资产组合的风险。

练习

方法

26. 某人想要建立一个资产组合。他考虑两只股票，令 x 表示股票1的投资收益率，y 表示股票2的投资收益率。股票1的期望收益和方差分别为 $E(x)=8.45\%$ 和 $\mathrm{Var}(x)=25$。股票2的期望收益和方差分别为 $E(y)=3.2\%$ 和 $\mathrm{Var}(y)=1$。收益的协方差 $\sigma_{xy}=-3$。
    a. 求股票1和股票2投资的标准差。用标准差度量风险，哪一只股票的投资风险更高？
    b. 某人用500美元投资购买股票1，求投资的期望收益和标准差。
    c. 某人建立一个资产组合，每只股票各投资50%，

求投资的期望收益和标准差（以美元计）。

d. 某人建立一个资产组合，70%投资于股票1，30%投资于股票，2，求投资的期望收益和标准差。

e. 计算x和y的相关系数，并评述这两只股票收益率之间的关系。

28. PortaCom开发设计了一款高品质的便携式打印机。直接人工和零件是制造成本的两个关键组成部分。在测试阶段，公司开发出产品的雏形并对新型打印机产品进行广泛的测试。PortaCom的工程师给出如下制造成本的二元概率分布。令随机变量x表示每台打印机的零件成本，随机变量y表示每台打印机的直接人工成本，单位是美元。管理者想要根据这个概率分布估计制造成本。

零件（x）	直接人工（y）			合计
	43	45	48	
85	0.05	0.2	0.2	0.45
95	0.25	0.2	0.1	0.55
合计	0.30	0.4	0.3	1.00

a. 给出直接人工成本的边际分布并计算其数学期望、方差和标准差。

b. 给出零件成本的边际分布并计算其数学期望、方差和标准差。

c. 单位制造总成本等于直接人工成本和零件成本之和。给出单位制造总成本的概率分布。

d. 计算单位制造总成本的数学期望、方差和标准差。

e. 直接人工成本和零件成本相互独立吗？为什么？如果你的结论是不独立，那么直接人工成本和零件成本之间的关系如何？

f. PortaCom生产了1500台打印机用于产品推介，总制造成本为198350美元。这和你预期的一样吗？如果比你预期的偏高或者偏低，你认为是什么原因？

30. 除了（29）题中S&P 500和核心债券的信息之外，摩根大通资产管理还报告了不动产投资信托公司（REITs）的期望收益率为13.07%，标准差为23.17%（摩根大通资产管理，《市场指南》，2012年第一季度）。S&P 500和REITs的相关系数为0.74；核心债券和REITs的相关系数为-0.04。

a. 根据所给出的信息，求S&P 500和核心债券的协方差。

b. 建立一个资产投资组合，S&P 500指数基金投资和REITs投资各占50%。求资产组合的期望收益率和标准差。

c. 建立一个资产投资组合，核心债券基金投资和REITs投资各占50%。求资产组合的期望收益率和标准差。

d. 建立一个资产投资组合，核心债券基金投资占80%，REITs投资占20%。求资产组合的期望收益率和标准差。

e. 你会建议一个激进的投资者采用（b）、（c）和（d）中哪一个资产投资组合？你会建议一个保守的投资者采用哪种资产投资组合？为什么？

5.5 二项概率分布

二项概率分布是一种离散型概率分布，具有广泛的应用。它与一个称为二项试验的多步骤试验有关。

5.5.1 二项试验

二项试验（binomial experiment）具有以下四个性质。

二项试验的性质

1. 试验由一系列相同的n个试验组成。
2. 每次试验有两种可能的结果。我们把其中一个称为成功，另一个称为失败。
3. 每次试验成功的概率都是相同的，用p来表示；失败的概率也都相同，用$1-p$表示。
4. 试验是相互独立的。

如果一个试验具有性质2、性质3和性质4，我们称该试验是由伯努利[⊖]过程产生的。另外，如果该试验还具

[⊖] 瑞士人雅各布·伯努利（Jakob Bernoulli，1654—1705年）是伯努利家族最杰出的数学家，发表了关于排列和组合理论以及二项式定理的概率论文。

有性质1，则称其为二项试验。图5-2给出了一个由成功和失败组成的序列，它是一个包含8次试验的二项试验。

在二项试验中，我们感兴趣的是在n次试验中成功出现的次数。如果令x代表n次试验中成功的次数，则x的可能值为0，1，2，3，…，n。由于随机变量取值的个数是有限的，所以x是一个离散型随机变量。与这一随机变量相对应的概率分布称为**二项概率分布**（binomial probability distribution）。例如，考虑抛5次硬币的试验，每一次都观察硬币着地时正面朝上还是反面朝上。假设我们想要计算5次抛掷中正面出现的次数。这个试验具备二项试验的性质吗？感兴趣的随机变量是什么？注意：

（1）试验由5次相同的试验组成，每次试验都是抛一枚硬币。

（2）每次试验都有两种可能的结果：正面朝上或反面朝上。定义正面朝上为成功，反面朝上为失败。

（3）在每次试验中，正面朝上和反面朝上的概率都是一样的，即$p=0.5$，$1-p=0.5$。

（4）因为任意一次试验的结果都不影响其他各次试验（抛掷），所以各次试验（抛掷）是独立的。

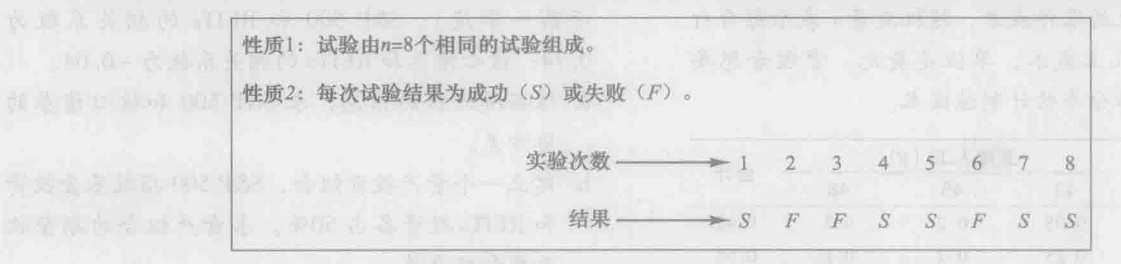


图5-2 在一个8次试验组成的二项试验中，由成功和失败组成的一个序列

于是，满足二项试验的性质。感兴趣的随机变量为x=抛掷5次硬币正面朝上的次数，这时x的可能取值为0，1，2，3，4或5。

考虑另一个例子，一名保险推销员随机选择10个家庭进行访问。每次访问的结果无外乎两种：成功——该户购买保险；失败——该户没有购买保险。根据过去的经验，推销员已知随机选择的家庭会购买保险的概率为0.10。它满足二项试验的性质吗？显然：

（1）试验由10次相同的试验组成，每一次试验是访问一个家庭。

（2）每次试验中有两种可能的结果：成功——该户购买保险；失败——该户没有购买保险。

（3）每次访问中，家庭是否购买保险的概率是相同的，即：$p=0.10$，$1-p=0.90$。

（4）由于家庭是随机选择的，所以试验是独立的。

这是一个二项试验，它满足二项试验的四条性质。感兴趣的随机变量是访问的10个家庭中购买保险的家庭数。这时，x的可能取值为0，1，2，3，4，5，6，7，8，9或10。

二项试验的性质3称为平稳性假设，它有时容易与性质4（试验的独立性）混淆。为了区分二者的不同，再次考虑推销员访问家庭推销保险的例子。如果一段时间后，推销员由于疲惫而失去热情，比如，到第10次访问时，成功（卖出保险）的概率下降到0.05。这时，性质3（平稳性）不能满足，从而试验不再是二项试验。即便该试验满足性质4（即每个家庭的购买决策是独立的），但如果不满足性质3，那么试验也不是一个二项试验。

在二项试验的应用中，常常用到一个称为二项概率函数的公式，它用于计算n次试验中x次成功的概率。利用第4章已介绍的有关概率的概念，下面我们通过例子说明如何使用这一公式。

5.5.2 马丁服装商店的问题

考虑光顾马丁服装商店的3名顾客购买服装的情况。根据过去的经验，商店经理估计每名顾客购买服装的概率是0.30。那么，在3名顾客中有2名顾客购买服装的概率是多少？

根据树形图（见图5-3），3名顾客中每名顾客各自作出购买决策，从而试验有8种可能的结果。用S表示成功（购买），F表示失败（未购买），我们感兴趣的是3次试验（购买决策）中有2次成功的试验结果。接下来，可以将3次购买决策组成的试验序列看作一个二项试验，它满足二项试验的四个基本条件：

(1) 可以将试验认为是由3个相同的试验组成的序列，在3名光临商店的顾客中，每名顾客对应于一次试验。

(2) 每次试验都只有两种试验结果：顾客购买（成功）或不购买（失败）。

(3) 假设对所有顾客而言，顾客购买的概率（0.30）或不购买的概率（0.70）都是相同。

(4) 每个顾客的购买决策与其他顾客的购买决策相互独立。

于是，满足二项试验的性质。

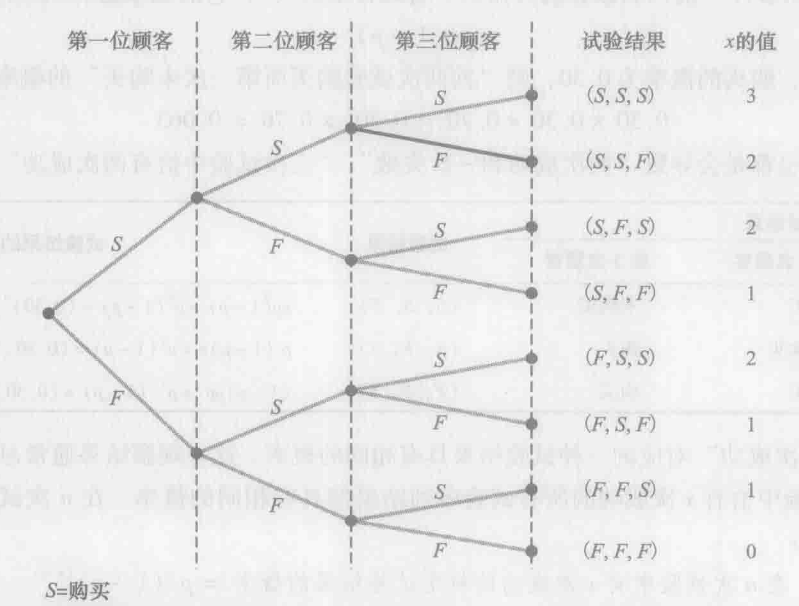


图5-3 马丁服装商店问题的树形图

在n次试验中，恰有x次成功的二项试验里，可用以下公式计算各种不同结果的数目：[⊖]

n次试验中恰有x次成功的试验结果的个数

$$\binom{n}{x}=\frac{n!}{x!(n-x)!} \tag{5-10}$$

式中，$n!=n(n-1)(n-2)\cdots(2)(1)$

并且定义

$$0!=1$$

现在仍回到马丁服装商店的试验中，它涉及3名顾客的购买决策。式（5-10）可用来确定“有2人购买”的试验结果的个数。即在$n=3$次试验中，有多少种方式能得到$x=$成功2次。由式（5-10）我们有：

$$\binom{n}{x}=\binom{3}{2}=\frac{3!}{2!(3-2)!}=\frac{3\times2\times1}{2\times1\times1}=\frac{6}{2}=3$$

式（5-10）表明，成功两次的试验结果有3种。从图5-3可见，这3种结果分别是（S, S, F），（S, F, S）和（F, S, S）。

利用式（5-10），可计算出3次试验中有3次成功（购买）的试验结果的个数为：

$$\binom{n}{x}=\binom{3}{3}=\frac{3!}{3!(3-3)!}=\frac{3!}{3!0!}=\frac{3\times2\times1}{3\times2\times1\times1}=\frac{6}{6}=1$$

⊖ 在第4章中，组合公式确定从n个对象中选择x个时有多少种不同的选法。对于二项试验，，组合公式计算出有多少种不同的试验结果与“n次试验序列中有x次成功”相对应。

从图5-3可见，只有一种试验结果是“3次成功”，记做（S，S，S）。

式（5-10）可用来确定与x次成功相对应的试验结果的个数。如果我们想要确定n次试验中x次成功的概率，还必须要知道其中每一个试验结果发生的概率。由于二项试验的各个试验是相互独立的，我们只需简单地将各个试验结果发生的概率相乘，就能得到某个由成功和失败所组成的特定试验序列发生的概率。

令（S，S，F）表示事件“前两名顾客购买而第3名顾客未购买”，它的概率是

$$pp(1-p)$$

由于在每次试验中，购买的概率为0.30，则“前两次试验购买而第三次未购买”的概率是

$$0.30\times0.30\times0.70=0.30^2\times0.70=0.063$$

还有两种试验结果也都是会导致“两次成功和一次失败”。“三次试验中恰有两次成功”的概率见下表：

试验结果			试验结果	试验结果的概率
第1名顾客	第2名顾客	第3名顾客		
购买	购买	未购买	(S, S, F)	$pp(1-p)=p^2(1-p)=(0.30)^2\times0.70=0.063$
购买	未购买	购买	(S, F, S)	$p(1-p)p=p^2(1-p)=(0.30)^2\times0.70=0.063$
未购买	购买	购买	(F, S, S)	$(1-p)pp=p^2(1-p)=(0.30)^2\times0.70=0.063$

可见，与“恰有两次成功”对应的三种试验结果具有相同的概率。这一观察结果通常总是成立的。在任何一个二项试验中，n次试验中恰有x次成功的所有试验序列结果都具有相同的概率。在n次试验中有x次成功的每个试验序列的概率如下：

$$\text{在 } n \text{ 次试验中有 } x \text{ 次成功的特定试验结果的概率}=p^x(1-p)^{(n-x)} \tag{5-11}$$

对于马丁服装商店问题，公式表明任何一个“两次成功”试验结果发生的概率为$p^2(1-p)^{(3-2)}=p^2(1-p)^1=(0.30)^2\times0.70=0.063$。

式（5-10）给出一个二项试验中有x次成功的试验结果的个数，式（5-11）给出了x次成功的每个试验序列的概率，结合式（5-10）和式（5-11）得到下面的**二项概率函数**（binomial probability function）：

二项概率函数

$$f(x)=\binom{n}{x}p^x(1-p)^{(n-x)} \tag{5-12}$$

式中，x为成功的次数；p为一次试验中成功的概率；n为试验的次数；$f(x)$为n次试验中有x次成功的概率；$\binom{n}{x}=\frac{n!}{x!(n-x)!}$。

对于二项概率分布，x是一个离散型随机变量，概率函数为$f(x)$，其中$x=0, 1, 2, 3, \cdots, n$。

以马丁服装商店问题为例，我们利用式（5-12）计算“没有顾客购买”的概率、“恰有一名顾客购买”的概率、“恰有两名顾客购买”的概率和“所有三名顾客均购买”的概率，如表5-6所示。计算结果汇总于表5-13中，它给出发生购物行为的顾客人数的概率分布。图5-4是它的概率分布图。

表5-13 购物的顾客人数的概率分布

x	$f(x)$	x	$f(x)$
0	$\frac{3!}{0!3!}(0.30)^0(0.70)^3=0.343$	2	$\frac{3!}{2!1!}(0.30)^2(0.70)^1=0.189$
1	$\frac{3!}{1!2!}(0.30)^1(0.70)^2=0.441$	3	$\frac{3!}{3!0!}(0.30)^3(0.70)^0=\underline{0.027}$

二项概率函数可用于任何二项试验。只要试验满足二项试验的性质，并且我们已知 n 和 p 的值，那么便可使用式（5-12）计算 n 次试验中有 x 次成功的概率。

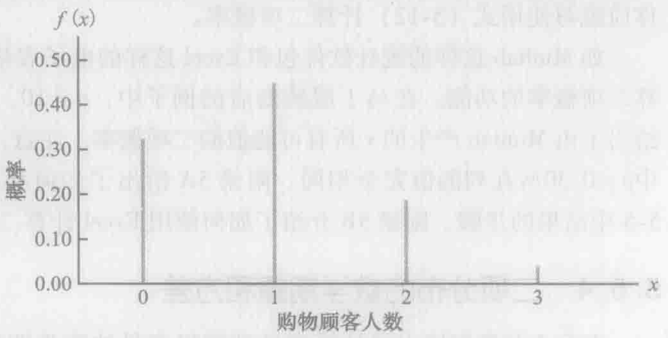


图 5-4　购物顾客人数的概率分布图

如果我们对马丁试验稍做变动，比如进入商店的顾客不是 3 名而是 10 名，由式（5-12）给出的二项概率函数仍是可用的。这是一个二项试验，其中 $n=10$，$x=4$，$p=0.30$。从而，10 名进入商店的顾客中恰好有 4 人购物的概率是

$$f(4)=\frac{10!}{4!6!}\times(0.30)^4\times(0.70)^6=0.2001$$

5.5.3　二项概率表的使用

人们已经编制出了数学用表，用于计算一个共有 n 次试验的二项试验中成功 x 次的概率。这些数学用表比式（5-12）使用起来更加容易和快捷。附录 B 中的表 5 是一个二项概率表，表 5-14 是从二项概率表中节选出的一部分。在使用二项概率表前，我们必须先确定二项试验中 n、p 和 x 的值。在表 5-14 的例子中，我们看到 $n=10$，$p=0.40$的二项试验中成功次数 $x=3$ 的概率为 0.215 0。如果你直接用二项概率函数式（5-12），也可以得到相同的答案。

表 5-14　从二项概率表查找概率值

例：$n=10$，$x=3$，$p=0.40$，$f(3)=0.2150$

n	x	p									
		0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
9	0	0.630 2	0.387 4	0.231 6	0.134 2	0.075 1	0.040 4	0.020 7	0.010 1	0.004 6	0.002 0
	1	0.298 5	0.387 4	0.367 9	0.302 0	0.225 3	0.155 6	0.100 4	0.060 5	0.033 9	0.017 6
	2	0.062 9	0.172 2	0.259 7	0.302 0	0.300 3	0.266 8	0.216 2	0.161 2	0.111 0	0.070 3
	3	0.007 7	0.044 6	0.106 9	0.176 2	0.233 6	0.266 8	0.271 6	0.250 8	0.211 9	0.164 1
	4	0.000 6	0.007 4	0.028 3	0.066 1	0.116 8	0.171 5	0.219 4	0.250 8	0.260 0	0.246 1
	5	0.000 0	0.000 8	0.005 0	0.016 5	0.038 9	0.073 5	0.118 1	0.167 2	0.212 8	0.246 1
	6	0.000 0	0.000 1	0.000 6	0.002 8	0.008 7	0.021 0	0.042 4	0.074 3	0.116 0	0.164 1
	7	0.000 0	0.000 0	0.000 0	0.000 3	0.001 2	0.003 9	0.009 8	0.021 2	0.040 7	0.070 3
	8	0.000 0	0.000 0	0.000 0	0.000 0	0.000 1	0.000 4	0.001 3	0.003 5	0.008 3	0.017 6
	9	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 1	0.000 3	0.000 8	0.002 0
10	0	0.598 7	0.348 7	0.196 9	0.107 4	0.056 3	0.028 2	0.013 5	0.006 0	0.002 5	0.001 0
	1	0.315 1	0.387 4	0.347 4	0.268 4	0.187 7	0.121 1	0.072 5	0.040 3	0.020 7	0.009 8
	2	0.074 6	0.193 7	0.275 9	0.302 0	0.281 6	0.233 5	0.175 7	0.120 9	0.076 3	0.043 9
	3	0.010 5	0.057 4	0.129 8	0.201 3	0.250 3	0.266 8	0.252 2	0.215 0	0.166 5	0.117 2
	4	0.001 0	0.011 2	0.040 1	0.088 1	0.146 0	0.200 1	0.237 7	0.250 8	0.238 4	0.205 1
	5	0.000 1	0.001 5	0.008 5	0.026 4	0.058 4	0.102 9	0.153 6	0.200 7	0.234 0	0.246 1
	6	0.000 0	0.000 1	0.001 2	0.005 5	0.016 2	0.036 8	0.068 9	0.111 5	0.159 6	0.205 1
	7	0.000 0	0.000 0	0.000 1	0.000 8	0.003 1	0.009 0	0.021 2	0.042 5	0.074 6	0.117 2
	8	0.000 0	0.000 0	0.000 0	0.000 1	0.000 4	0.001 4	0.004 3	0.010 6	0.022 9	0.043 9
	9	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 1	0.000 5	0.001 6	0.004 2	0.009 8
	10	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 1	0.000 3	0.001 0

现在，我们使用表 5-14 求解马丁服装商店问题，计算 10 次试验中有 4 次成功的概率。此时，$n=10$，$x=4$，$p=0.30$，直接查二项概率表得到$f(4)=0.2001$。

尽管二项概率表易于使用，但表中不可能列出二项试验可能遇到的所有 n 和 p 值。如今根据式（5-12），我们使

用计算器来计算想要的概率并不困难，尤其在试验的数目不大时。在习题中，除非题目特别要求你使用二项概率表，你应练习使用式（5-12）计算二项概率。

如 Minitab 这样的统计软件包和 Excel 这样的电子表格软件包都具有计算二项概率的功能。在马丁服装商店的例子中，$n=10$，$p=0.30$。图5-5给出了由 Minitab 产生的 x 所有可能值的二项概率。注意，这些值与表5-14中 $p=0.30$ 所在列的值完全相同。附录 5A 给出了如何利用 Minitab 得到图5-5 中结果的步骤。附录 5B 介绍了如何使用 Excel 计算二项概率。

x	$P(X=x)$
0	0.0282
1	0.1211
2	0.2335
3	0.2668
4	0.2001
5	0.1029
6	0.0368
7	0.0090
8	0.0014
9	0.0001
10	0.0000

图 5-5 Minitab 输出结果——马丁服装店问题中的二项概率

5.5.4 二项分布的数学期望和方差

在 5.3 节我们给出了计算离散型随机变量的数学期望和方差的公式。在随机变量服从二项分布的特定情形下，已知试验次数为 n，成功的概率为 p，数学期望和方差公式得以简化，具体形式如下。

二项分布的数学期望与方差

$$E(x) = \mu = np \tag{5-13}$$

$$\mathrm{Var}(x) = \sigma^2 = np(1-p) \tag{5-14}$$

在马丁服装店的例子中，当有 3 名顾客光顾时，利用式（5-13）计算购物顾客人数的数学期望，有

$$E(x) = np = 3 \times 0.30 = 0.9$$

假定预计下个月马丁服装商店有 1 000 名顾客光顾，那么购物顾客人数的期望值是多少？答案是 $\mu=np=1\,000\times 0.3=300$。因此，为增加购物人数的期望值，马丁服装商店必须吸引更多顾客光顾并且（或者）提高每名到店顾客购物的概率。

在马丁服装店的例子中，当有 3 名顾客光顾时，购物人数的方差和标准差为：

$$\sigma^2 = np(1-p) = 3 \times 0.3 \times 0.7 = 0.63$$

$$\sigma = \sqrt{0.63} = 0.79$$

在随后光临的 1 000 名顾客中，购物人数的方差和标准差为：

$$\sigma^2 = np(1-p) = 1\,000 \times 0.3 \times 0.7 = 210$$

$$\sigma = \sqrt{210} = 14.49$$

注释和评论

1. 附录 B 中的二项概率表给出了 $p\leqslant 0.95$ 的概率值。有的二项概率表中，可能仅给出了 $p\leqslant 0.50$ 的概率值。当成功的概率 $p>0.50$ 时，这样的表似乎是不可用的。然而，由于 $n-x$ 次失败的概率也是 x 次成功的概率，因此这样的表仍然是可用的。当成功的概率 $p>0.50$ 时，我们可以转而计算 $n-x$ 次失败的概率。当 $p>0.50$时，失败的概率 $1-p\leqslant 0.50$。
2. 某些二项概率表是以累计形式给出的。在使用这种表计算 n 次试验中恰好有 x 次成功的概率时，必须减去上一项的值。例如，$f(2)=P(x\leqslant 2)-P(x\leqslant 1)$。附录 B 中的表格直接给出的是概率值 $f(2)$。当使用附录 B 中的表计算累计概率时，只需将表中的概率值相加即可。例如，累计概率 $P(x\leqslant 2)=f(0)+f(1)+f(2)$。

练 习

方法

32. 考虑一个二项试验，其中 $n=10$，$p=0.10$。
    a. 计算 $f(0)$。
    b. 计算 $f(2)$。
    c. 计算 $P(x\leqslant 2)$。
    d. 计算 $P(x\geqslant 1)$。

e. 计算 $E(x)$。

f. 计算 $\text{Var}(x)$ 和 σ。

应用

34. 对洲际酒店及度假村进行一项哈里斯互动调查，询问受访者："在国际旅行中，你通常是为了文化体验而独自冒险还是随旅行团按行程出行？"调查发现，有23%的受访者选择跟随旅行团（《今日美国》，2004 年 1 月 21 日）。

a. 在一个由 6 名国际游客组成的样本中，恰好有两人跟随旅行团旅行的概率是多少？

b. 在一个由 6 名国际游客组成的样本中，至少有两人跟随旅行团旅行的概率是多少？

c. 在一个由10 名国际游客组成的样本中，没有人随旅行团旅行的概率是多少？

☆36. 当新机器正常运转时，只有3%的产品是废品。假设随机选取机器生产的两个零部件，我们感兴趣的是其中废品的件数。

a. 若这个试验是二项试验，应满足哪些条件？

b. 绘制与图 5-3 类似的树形图，说明这是一个由两次试验组成的试验序列。

c. 有多少个试验结果会导致"恰好发现一个废品"？

d. 计算"没有发现废品"、"恰好发现一件废品"以及"恰好发现两件废品"的概率。

38. 在敌人入侵时，设计要求军事雷达和导弹探测系统发出警报。可靠性指的是探测系统是否能发现攻击并发出警告。假设某探测系统能探测到导弹攻击的概率为 0.90，利用二项概率分布回答以下问题。

a. 单个探测系统能够发现攻击的概率是多少？

b. 如果将两套探测系统安装在同一地区独立工作，则至少有一套探测系统能发现攻击的概率是多少？

c. 如果安装了三套探测系统，则至少有一套系统能发现攻击的概率是多少？

d. 你建议使用多套探测系统吗？为什么？

40. 美国人口普查局最近的人口调查显示，年龄在 25 岁及以上的人中有 28% 已经完成了四年制的大学学业（《纽约时报年鉴》，2006 年）。对于由15 名年龄在 25 岁及以上的人组成的一个样本，回答下列问题：

a. 恰好有 4 人完成了四年制的大学学业的概率是多少？

b. 至少有 3 人完成了四年制的大学学业的概率是多少？

42. 据亚美利交易控股公司进行的一次调查发现，有 1/4 的投资者在其资产投资组合中包括有交易型开放式指数基金（《今日美国》，2007 年 1 月 11 日）。对于由20 名投资者组成的一个样本，求：

a. 恰好有 4 名投资者在其投资组合中有交易型开放式指数基金的概率。

b. 至少有 2 名投资者在其投资组合中有交易型开放式指数基金的概率。

c. 如果你发现恰有 12 名投资者在其投资组合中有交易型开放式指数基金，你会对调查结果的精确度产生怀疑吗？

d. 计算在投资组合中有交易型开放式指数基金的投资者人数的数学期望。

5.6 泊松概率分布

本节我们介绍一个常用的离散型随机变量，它主要用于估计在特定时间段或空间中某事件发生的次数。例如，我们感兴趣的随机变量可能是一小时内到达洗车房的汽车数、10 英里长的高速公路上需要维修的路段数目或者 100 英里长的水管有多少处发生泄漏。如果事件出现的次数满足以下两个性质，则随机变量服从**泊松概率分布**（poisson probability distribution）[⊖]。

泊松试验的性质

1. 在任意两个相等长度的区间上，事件发生的概率相等。
2. 事件在某一区间上是否发生与事件在其他区间上是否发生是独立的。

⊖ 泊松概率分布经常用来模拟排队时随机到达的数目。

泊松概率函数（poisson probability function）由式（5-15）给出：

泊松概率函数[㊀]

$$f(x)=\frac{\mu^{x}e^{-\mu}}{x!} \tag{5-15}$$

式中，$f(x)$ 为事件在一个区间发生 x 次的概率；μ 为事件在一个区间发生次数的数学期望或均值；$e=2.718\ 28$

对泊松概率分布，x 是一个离散型随机变量，它表示区间上事件发生的次数。由于没有对事件发生的次数指明上限，因此在概率函数 $f(x)$ 中，有 $x=0$，1，2，…即随机变量的取值有无限多种可能。在实际应用中，当 x 最终取值非常大时，$f(x)$ 近似为0。因此，对一些太大的 x 值，可忽略其发生的可能性。

5.6.1 一个时间段上的例子

假定我们感兴趣的是，工作日早上15分钟内到达某汽车银行出纳窗口处的汽车数量。如果假设在任意两个相等长度的时间段内汽车到达的概率是相等的，并且在任意一段时间内是否有汽车到达与其他时段内是否有汽车到达是相互独立的，那么泊松概率函数[㊁]是适用的。假定以上假设都成立，并且对历史数据的分析显示，15分钟的时间段中到达的车辆数目的平均值为10。这时，采用以下概率函数：

$$f(x)=\frac{10^{x}e^{-10}}{x!}$$

其中，随机变量 $x=15$ 分钟的时间段中到达的汽车数。

如果管理人员想要知道15分钟内恰有5辆汽车到达的概率，那么令 $x=5$，有：

$$15\text{分钟内恰有5辆汽车到达的概率}=f(5)=\frac{10^{5}e^{-10}}{5!}=0.037\ 8$$

尽管在 $\mu=10$，$x=5$ 处计算概率函数就可以确定这一概率，但通常查泊松分布表更为方便。泊松分布表给出了与特定的 x 和 μ 相对应的概率，参见附录B中表7。为使用方便，表5-15是从表7中节选的一部分。注意，查表时我们只需知道 x 和 μ 的值即可使用泊松概率表。从表5-15可见，位于表中 $x=5$ 的行和 $\mu=10$ 的列交叉位置上的值就是15分钟内有5辆汽车到达的概率。从而，我们得到 $f(5)=0.037\ 8$。

表5-15 从泊松概率表查找概率值概率表节选值
例：$\mu=10$，$x=5$，$f(5)=0.037\ 8$

x	μ									
	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10
0	0.000 1	0.000 1	0.000 1	0.000 1	0.000 1	0.000 1	0.000 1	0.000 1	0.000 1	0.000 0
1	0.001 0	0.000 9	0.000 9	0.000 8	0.000 7	0.000 7	0.000 6	0.000 5	0.000 5	0.000 5
2	0.004 6	0.004 3	0.004 0	0.003 7	0.003 4	0.003 1	0.002 9	0.002 7	0.002 5	0.002 3
3	0.014 0	0.013 1	0.012 3	0.011 5	0.010 7	0.010 0	0.009 3	0.008 7	0.008 1	0.007 6
4	0.031 9	0.030 2	0.028 5	0.026 9	0.025 4	0.024 0	0.022 6	0.021 3	0.020 1	0.018 9
5	0.058 1	0.055 5	0.053 0	0.050 6	0.048 3	0.046 0	0.043 9	0.041 8	0.039 8	0.037 8
6	0.088 1	0.085 1	0.082 2	0.079 3	0.076 4	0.073 6	0.070 9	0.068 2	0.065 6	0.063 1
7	0.114 5	0.111 8	0.109 1	0.106 4	0.103 7	0.101 0	0.098 2	0.095 5	0.092 8	0.090 1
8	0.130 2	0.128 6	0.126 9	0.125 1	0.123 2	0.121 2	0.119 1	0.117 0	0.114 8	0.112 6
9	0.131 7	0.131 5	0.131 1	0.130 6	0.130 0	0.129 3	0.128 4	0.127 4	0.126 3	0.125 1

㊀ 西蒙·泊松于1802～1808年在巴黎Ecole Polytechnique讲授数学。在他1837年发表的题目为"*Researches on the Probability of Criminal and Civil Verdicts*"的文章中，对被后人称为泊松分布的分布进行了研究。

㊁ 贝尔实验室利用泊松分布模拟接到的电话次数。

（续）

x	μ									
	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10
10	0.119 8	0.121 0	0.121 9	0.122 8	0.123 5	0.124 1	0.124 5	0.124 9	0.125 0	0.125 1
11	0.099 1	0.101 2	0.103 1	0.104 9	0.106 7	0.108 3	0.109 8	0.111 2	0.112 5	0.113 7
12	0.075 2	0.077 6	0.079 9	0.082 2	0.084 4	0.086 6	0.088 8	0.090 8	0.092 8	0.094 8
13	0.052 6	0.054 9	0.057 2	0.059 4	0.061 7	0.064 0	0.066 2	0.068 5	0.070 7	0.072 9
14	0.034 2	0.036 1	0.038 0	0.039 9	0.041 9	0.043 9	0.045 9	0.047 9	0.050 0	0.052 1
15	0.020 8	0.022 1	0.023 5	0.025 0	0.026 5	0.028 1	0.029 7	0.031 3	0.033 0	0.034 7
16	0.011 8	0.012 7	0.013 7	0.014 7	0.015 7	0.016 8	0.018 0	0.019 2	0.020 4	0.021 7
17	0.006 3	0.006 9	0.007 5	0.008 1	0.008 8	0.009 5	0.010 3	0.011 1	0.011 9	0.012 8
18	0.003 2	0.003 5	0.003 9	0.004 2	0.004 6	0.005 1	0.005 5	0.006 0	0.006 5	0.007 1
19	0.001 5	0.001 7	0.001 9	0.002 1	0.002 3	0.002 6	0.002 8	0.003 1	0.003 4	0.003 7
20	0.000 7	0.000 8	0.000 9	0.001 0	0.001 1	0.001 2	0.001 4	0.001 5	0.001 7	0.001 9
21	0.000 3	0.000 3	0.000 4	0.000 4	0.000 5	0.000 6	0.000 6	0.000 7	0.000 8	0.000 9
22	0.000 1	0.000 1	0.000 2	0.000 2	0.000 2	0.000 2	0.000 3	0.000 3	0.000 4	0.000 4
23	0.000 0	0.000 1	0.000 1	0.000 1	0.000 1	0.000 1	0.000 1	0.000 1	0.000 2	0.000 2
24	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 1	0.000 1	0.000 1

在上面的例子中，泊松分布的数学期望$\mu=10$，即15分钟内平均到达10辆车。泊松分布的一个重要性质是它的数学期望和方差相等。从而，15分钟内到达车辆数目的方差$\sigma^2=10$，标准差$\sigma=\sqrt{10}=3.16$。

我们的例子中时间长度是15分钟，这种方法对其他长度的时间段也同样适用。假定我们要计算3分钟内有1辆汽车到达的概率。由15分钟内到达车辆的期望数为10可得，1分钟内到达的车辆的期望数为10/15＝2/3，于是3分钟内到达车辆的期望数为3×2/3＝2。令$\mu=2$，由下面的泊松概率函数即可求得3分钟内到达x辆车的概率：

$$f(x)=\frac{2^x e^{-2}}{x!}$$

3分钟内有一辆车到达的概率计算如下：

$$3\text{ 分钟内恰有 1 辆汽车到达的概率}=f(1)=\frac{2^1 e^{-2}}{1!}=0.2707$$

在前面的例子中，15分钟内恰有5辆汽车到达的概率是0.037 8。与3分钟内恰有1辆汽车到达的概率（0.270 7）是不同的。因此，在计算不同长度的时间段上泊松概率时，必须先计算在相应区间上随机变量的平均到达比率，然后再计算其概率。

5.6.2 一个长度或距离区间上的例子

下面，我们介绍一个不是在一段时间区间上应用泊松分布的例子。假定我们关心的是高速公路在重新整修后的一个月内出现严重损坏地段的数目。假设任意两段相等长度的高速公路上出现一处损坏的概率是相等的，并且任意一段距离上是否出现损坏与另一段距离上是否出现损坏无关。因此，泊松分布是适用的。

假定我们知道在重新整修后的一个月内，平均每英里有两处受到严重损坏。求某段3英里长的高速公路上没有发生严重损坏的概率。由于我们感兴趣的距离长度是3英里，显然3英里长的高速公路上出现严重损坏地段的期望数$\mu=2$处/英里×3英里＝6处。根据式（5-15）可得，没有严重损坏的概率为$f(0)=6^0\times e^{-6}/0!=0.0025$。因此，3英里长的距离没有发生严重损坏的情形几乎不可能发生。事实上，这个例子说明这段路至少发生1处严重损坏的概率为1－0.002 5＝0.997 5。

练习

方法

44. 考虑一个泊松分布，其中$\mu=3$。
  a. 写出相应的泊松概率函数。
  b. 计算$f(2)$。
  c. 计算$f(1)$。
  d. 计算$P(x\geqslant2)$。

应用

46. Regional 航空公司的预订票处平均每小时接到 48 个电话。
  a. 计算 5 分钟内接到 3 个电话的概率。
  b. 计算 15 分钟内恰好接到 10 个电话的概率。
  c. 假设现在没有等待电话。如果代理人需花费 5 分钟时间才能结束目前的通话，那么在这段时间里预期有多少个等待电话？没有等待电话的概率是多少？
  d. 如果现在没有电话，接下来代理人可休息 3 分钟而不被打扰的概率是多少？

☆48. 去年，含早餐服务的旅馆（B&Bs）接待的客人超过了 5 000 万。平均每分钟大约有七个人登录北美洲含早餐服务的旅馆的网址。通过网络，许多 B&Bs 得以招揽客人（《时代周刊》，2001 年 9 月）。
  a. 计算 1 分钟的时段中无人登录该网站的概率。
  b. 计算 1 分钟的时段中至少有两人登录该网站的概率。
  c. 计算30 秒的时段中至少有一人登录该网站的概率。
  d. 计算 1 分钟的时段中至少有 5 人登录该网站的概率。

50. 每年飞行事故的平均次数为 15 次（*The World Almanac and Book of Facts*，2004 年）。
  a. 计算每月飞行事故的平均值。
  b. 求 1 个月中没有发生飞行事故的概率。
  c. 求 1 个月中恰好发生 1 次飞行事故的概率。
  d. 求 1 个月中发生飞行事故多于 1 次的概率。

5.7 超几何概率分布

超几何概率分布（hypergeometric probability distribution）与二项概率分布联系密切。这两种概率分布主要有两处不同：超几何分布中的各次试验不是独立的；各次试验中成功的概率不等。

对于超几何分布中，符号 N 表示总体容量，r 表示总体中具有成功标志的元素的个数，$N-r$ 表示总体中具有失败标志的元素的个数。采用不放回抽样方法，从总体中抽取 n 个元素，**超几何概率函数**（hypergeometric probability function）用来计算在这 n 个元素中恰有 x 个元素具有成功标志，$n-x$ 个元素具有失败标志的概率。当这种试验结果出现时，我们是从总体的 r 个具有成功标志的元素中抽取 x 个，从总体的 $N-r$ 个具有失败标志的元素中抽取 $n-x$ 个。下面的超几何概率函数 $f(x)$ 给出了 n 次试验中有 x 次成功的概率。

超几何概率函数

$$f(x)=\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \tag{5-16}$$

式中，x 为成功的次数；n 为试验次数；$f(x)$ 为 n 次试验中 x 次成功的概率；N 为总体中元素个数；r 为总体中具有成功标志的元素的个数。

注意，$\binom{N}{n}$给出从一个容量为 N 的总体中抽取 n 个元素有多少种不同的抽取方式；$\binom{r}{x}$表示从总体 r 个具有成功标志的元素中抽取 x 个元素有多少种不同的抽取方式；$\binom{N-r}{n-x}$表示从总体 $N-r$ 个具有失败标志的元素中抽取 $n-x$个元素有多少种不同的抽取方式。

对于超几何概率分布，x 是一个离散型随机变量，在式（5-16）给出的概率函数$f(x)$ 中 $x=0, 1, 2, 3, \cdots, n$。

但是，只有观测到的成功次数小于或等于总体中的成功次数（$x \leqslant r$），并且观测到的失败次数小于或等于总体中的失败次数（$n-x \leqslant N-r$）时，x 的取值才是有效的。对一个或多个 x 值，如果这两个条件不成立，则 $f(x)=0$，即 x 取这些值的概率为0。

为说明如何利用式（5-16），我们考虑其在质量控制中的应用。安大略电器公司生产的保险丝每盒有12件。假定一名质检人员从一盒产品中随机选取3件进行检验。若这盒产品中恰好有5个保险丝是次品，则质检员抽出的3件产品中恰好发现1件次品的概率是多少？此时，$n=3$，$N=12$，$r=5$，$x=1$，从而

$$f(1)=\frac{\binom{5}{1}\binom{7}{2}}{\binom{12}{3}}=\frac{\left(\frac{5!}{1!4!}\right)\left(\frac{7!}{2!5!}\right)}{\left(\frac{12!}{3!9!}\right)}=\frac{5\times 21}{220}=0.4773$$

现在假定我们想知道至少有1件次品的概率。那么，一种最简单的做法是先计算出质检员没有发现次品的概率。$x=0$ 的概率是

$$f(0)=\frac{\binom{5}{0}\binom{7}{3}}{\binom{12}{3}}=\frac{\left(\frac{5!}{0!5!}\right)\left(\frac{7!}{3!4!}\right)}{\left(\frac{12!}{3!9!}\right)}=\frac{1\times 35}{220}=0.1591$$

由于0件次品的概率 $f(0)=0.1591$，故得到结论认为：至少有1件次品的概率是 $1-0.1591=0.8409$。因此，质检员至少发现1件次品的概率是相当高的。

超几何分布的均值和方差如下。

$$E(x)=\mu=n\left(\frac{r}{N}\right) \tag{5-17}$$

$$\operatorname{Var}(x)=\sigma^2=n\left(\frac{r}{N}\right)\left(1-\frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) \tag{5-18}$$

在前面的例子中，$n=3$，$r=5$，$N=12$，于是次品保险丝件数的均值和方差是

$$\mu=n\left(\frac{r}{N}\right)=3\times\frac{5}{12}=1.25$$

$$\sigma^2=n\left(\frac{r}{N}\right)\left(1-\frac{r}{N}\right)\left(\frac{N-n}{N-1}\right)=3\times\frac{5}{12}\times\left(1-\frac{5}{12}\right)\times\frac{12-3}{12-1}=0.60$$

标准差 $\sigma=\sqrt{0.60}=0.77$。

注释和评论

考虑 n 次试验的超几何分布，令 $p=r/N$ 表示首次试验中成功的概率。当总体容量足够大的时候，式（5-18）中的 $(N-n)/(N-1)$ 近似等于1。于是，数学期望 $E(x)=np$，方差 $\operatorname{Var}(x)=np(1-p)$。这与二项分布的数学期望和方差的表达式相同，见式（5-13）和式（5-14）。因此，当总体容量足够大的时候，超几何分布可以用试验次数为 n，成功概率 $p=r/N$ 的二项分布近似。

练习

方法

☆52. 假定 $N=10$，$r=3$。计算 n 和 x 取下列值时的超几何概率。

a. $n=4$，$x=1$。 b. $n=2$，$x=2$。

c. $n=2$，$x=0$。 d. $n=4$，$x=2$。

e. $n=4$，$x=4$。

应用

54. 在盖洛普进行的一项调查中，受访者被问及“你最喜欢观看的运动是什么？”按照偏爱程度分，足球和篮球分别排名第一和第二（Gallupwebsite，

2004年1月3日)。假设一个10人小组中，有7人偏爱足球，有3人偏爱篮球。从小组中抽出3个人组成一个随机样本。求：

a. 恰有2人偏爱足球的概率是多少？

b. 多数人（2人或3人）偏爱足球的概率是多少？

☆56. Axline计算机公司在两个工厂生产个人电脑，一个位于得克萨斯，另一个位于夏威夷。得克萨斯的工厂有40名员工，夏威夷的工厂有20名员工。随机选取10名员工组成一个样本，要求他们填写福利调查问卷。

a. 样本中没有夏威夷工厂员工的概率是多少？

b. 样本中有1名夏威夷工厂员工的概率是多少？

c. 样本中至少有2名夏威夷工厂员工的概率是多少？

d. 样本中有9名得克萨斯工厂员工的概率是多少？

58. 2008年10月美国国会通过了问题资产救助计划(TARP)，为苦苦挣扎的美国经济提供了70亿美元的帮助。其中超过20亿美元拨给了陷入困境的金融机构，希望能增加贷款助推经济复苏。但是，3个月过去了，美联储经调查发现，接受TARP基金的银行中有2/3收紧了商业贷款项目（《华尔街日报》，2009年2月3日）。在接受TARP基金最多的10家银行中，实际上这期间仅有3家银行增加了贷款。

增加贷款	减少贷款
美国BB&T公司	美国银行
太阳信托银行	美国第一资本金融公司
合众银行	花旗银行
	五三银行
	摩根大通
	地区金融公司
	美国富国银行

本题的目的是，假定你随机地从这10家银行中选取3家银行进行研究，以便继续对银行贷款业务进行监督。令随机变量x表示研究中增加了贷款的银行数。

a. $f(0)$ 是多少？你对该值做何解释？

b. $f(3)$ 是多少？你对该值做何解释？

c. 计算$f(1)$ 和$f(2)$。求研究中增加了贷款的银行数的概率分布。x取多大值时，概率最大？

d. 至少有一家银行增加了贷款的概率是多少？

e. 计算随机变量的数学期望、方差和标准差。

小 结

随机变量是对试验结果的数值描述。随机变量的概率分布描述了随机变量取不同值的概率。对任何离散型随机变量x，可以通过概率函数来定义概率分布，记做$f(x)$。它给出了随机变量取每一个值的概率。

我们介绍了离散型概率分布的两类形式。一类是在表格中列出随机变量的值和相应的概率。我们展示了如何将分配概率的相对频率法用于建立经验离散型概率分布。我们还讨论了二元经验分布。对于二元分布，兴趣的焦点是两个随机变量的关系。我们展示了如何计算协方差和相关系数度量二者的关系。我们还展示了金融资产市场收益的二元分布如何用于构建金融资产组合。

我们所讨论的第二类离散型概率分布是用数学函数表示随机变量的概率。二项分布、泊松分布和超几何分布都属于这种类型。当试验满足如下性质时，可使用二项分布确定n次试验中有x次成功的概率。

1. 试验由一系列完全相同的n个试验组成。

2. 每次试验有两种可能结果。一个称为成功，另一个称为失败。

3. 成功的概率p不随试验不同而不同，从而失败的概率$1-p$也不随试验不同而不同。

4. 试验是相互独立的。

当以上4条性质都满足时，可使用二项概率函数确定n次试验有x次成功的概率。我们还给出了计算二项分布的均值和方差的公式。

泊松分布用于确定事件在特定时间段或空间中发生x次的概率。在应用泊松分布时，必须满足以下假设：

1. 在任意两个相等长度的区间上事件发生的概率是相等的。

2. 事件在某一区间上是否发生与其他区间上事件是否发生独立。

在5.7节中介绍了第三种离散型概率分布——超几何概率分布。像二项分布一样，它也被用来计算n次试验中有x次成功的概率，但与二项分布不同的是，各次试验中成功的概率不同。

关键术语

random variable **随机变量** 对试验结果的数值描述。

discrete random variable **离散型随机变量** 有限多个取值或可数无限多个取值的随机变量。

continuous random variable **连续型随机变量** 在某一区间或多个区间内任意取值的随机变量。

probability distribution **概率分布** 对于随机变量的可能取值，描述概率如何分布。

probability function **概率函数** 是一个记做$f(x)$的函数，该函数给出了离散型随机变量取特定值x的概率。

empirical discrete distribution **经验离散分布** 一种概率分布，使用相对频率的方法分配概率

discrete uniform probability distribution **离散型均匀概率分布** 一种概率分布，随机变量取每个值的概率相等。

expected value **数学期望** 对随机变量中心位置的一种度量。

variance **方差** 对随机变量的变异性或离散性的一种度量。

standard deviation **标准差** 方差的算术平方根。

bivariate probability distribution **二元概率分布** 关于两个随机变量的概率分布。离散型二元概率分布给出两个随机变量的每对可能取值的概率。

binomial experiment **二项试验** 具有第5.5节中四个性质的试验。

binomial probability distribution **二项概率分布** 一种概率分布，给出二项试验中n次试验有x次成功的概率。

binomial probability function **二项概率函数** 用来计算二项概率的函数。

poisson probability distribution **泊松概率分布** 一种概率分布，给出在一段特定时间或空间中某个事件发生x次的概率。

poisson probability function **泊松概率函数** 用来计算泊松概率的函数。

hypergeometric probability distribution **超几何概率分布** 是一种概率分布，当从具有r个“成功”元素和$N-r$个“失败”元素的总体中抽取n次时，给出恰好有x次成功的概率。

hypergeometric probability function **超几何概率函数** 用来计算超几何概率的函数。

重要公式

离散型均匀概率函数

$$f(x)=1/n \tag{5-3}$$

离散型随机变量的数学期望

$$E(x)=\mu=\sum xf(x) \tag{5-4}$$

离散型随机变量的方差

$$\mathrm{Var}(x)=\sigma^2=\sum(x-\mu)^2f(x) \tag{5-5}$$

随机变量x和y的协方差

$$\sigma_{xy}=[\mathrm{Var}(x+y)-\mathrm{Var}(x)-\mathrm{Var}(y)]/2 \tag{5-6}$$

随机变量x和y的相关系数

$$\rho_{xy}=\frac{\sigma_{xy}}{\sigma_x\sigma_y} \tag{5-7}$$

随机变量x和y的线性组合的数学期望

$$E(ax+by)=aE(x)+bE(y) \tag{5-8}$$

两个随机变量的线性组合的方差

$$\mathrm{Var}(ax+by)=a^2\mathrm{Var}(x)+b^2\mathrm{Var}(y)+2ab\sigma_{xy} \tag{5-9}$$

式中，σ_{xy}是x和y的协方差。

n次试验中恰有x次成功时，试验结果的数目

$$\binom{n}{x}=\frac{n!}{x!(n-x)!} \tag{5-10}$$

二项概率函数

$$f(x)=\binom{n}{x}p^x(1-p)^{(n-x)} \tag{5-12}$$

二项分布的数学期望

$$E(x)=\mu=np \tag{5-13}$$

二项分布的方差

$$\mathrm{Var}(x)=\sigma^2=np(1-p) \tag{5-14}$$

泊松概率函数

$$f(x)=\frac{\mu^x e^{-\mu}}{x!} \tag{5-15}$$

超几何概率函数

$$f(x)=\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \tag{5-16}$$

超几何分布的数学期望

$$E(x) = \mu = n\left(\frac{r}{N}\right) \quad (5\text{-}17)$$

超几何分布的方差

$$\mathrm{Var}(x) = \sigma^2 = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right) \quad (5\text{-}18)$$

补充练习

60. 美国个人投资者协会向高级互助基金公布了一份年度指南（《高级互助基金个人投资指南》，22*e*，美国个人投资者协会，2003 年）。下表是互助基金中 29 类项目总的风险等级。

风险	基金数目
低	7
低于平均水平	6
平均水平	3
高于平均水平	6
高	7

a. 令 $x=1$ 表示低风险，$x=5$ 表示高风险。求风险水平的概率分布。

b. 求总风险的数学期望和方差。

c. 有 11 项基金属于债券基金。在债券基金中，有 7 类基金属于低风险，4 类基金风险属于低于平均水平。对比分析 18 只股票基金及债券基金的总风险。

62. 位于亚特兰大市哈茨菲尔德－杰克逊机场的书亭出售报刊（平装书、报纸和杂志）和零食（花生、椒盐卷饼和糖果等）。销售点终端（POS）中收集了消费者大量的购买信息。下表中给出了最近 600 名顾客购买零食和报刊的数量。

		阅读材料		
		0	1	2
	0	0	60	18
零食	1	240	90	30
	2	120	30	12

a. 随机选取一名消费者，令 x = 零食购买量，y = 报刊购买量。根据表中数据，求 x 和 y 的经验离散型二元概率分布。求顾客购买 1 种报刊和 2 种零食的概率。求顾客只购买 1 种零食的概率。概率 $f(x=0, y=0) = 0$，为什么？

b. 给出零食购买量的边际概率分布，并求其数学期望和标准差。

c. 求报刊购买量的数学期望和标准差。

d. 令 t = 零食和报刊购买量之和，给出 t 的概率分布，并求其数学期望和标准差。

e. 计算 x 和 y 的协方差和相关系数。顾客购买零食数量和报刊数量之间的关系如何？

64. 调查发现，使用月票的人单程从家到工作单位均需要耗时 26 分钟。此外，有 50% 的使用月票的人从家到工作单位单程耗时超过 1 小时。（交通统计署网站，2004 年 1 月）。

a. 如果某天调查 20 名使用月票的人，则有 3 人单程耗时超过 1 小时的概率是多少？

b. 如果某天调查 20 名使用月票的人，则没有人单程耗时超过 1 小时的概率是多少？

c. 如果某公司有 2 000 名雇员，则从家到工作单位单程耗时超过 1 小时的使用月票的雇员人数的期望值是多少？

d. 如果某公司有 2 000 名雇员，则从家到工作单位单程耗时超过 1 小时的使用月票的雇员人数的方差和标准差是多少？

66. 许多公司使用一种叫作“可接受抽样”的质量控制技术检测运来的零部件、原材料等。在电子行业中，通常零部件都是从供应商那里大批量购进。检测一个零件样本，可看作一个包含 n 次试验的二项试验。每一次零件检测（试验）的结果为零件完好或损坏。假如一批零件中损坏率不超过 1%，那么 Reynolds 电子公司就会接受供应商的这批零件。假设从新近运来的一批零件中随机抽出 5 件组成样本进行检测。

a. 假设这批零件有 1% 损坏，求样本中没有零件损坏的概率。

b. 假设这批零件有 1% 损坏，求样本中恰有 1 件损坏的概率。

c. 如果这批零件有 1% 损坏，求样本中至少有 1 件损坏的概率。

d. 如果发现有 1 件零件被损坏，这批货物你会欣然接受吗？为什么？

68. 佐格比国际公司的一项民意调查显示，在那些认为音乐在他们的生活中发挥了“非常重要的作用”的美国人中，有 30% 的受访者表示当地的广播电

台“总是”播放他们喜欢的音乐，（Zogby 网站，2004 年 1 月 12 日）。假定由 800 名认为音乐在他们的生活中发挥了“非常重要的作用”的美国人组成一个样本。

a. 你预期有多少人认为当地的广播电台“总是”播放他们喜欢的音乐？

b. 认为当地的广播电台“总是”播放他们喜欢的音乐的人数的标准差是多少？

c. 认为当地的广播电台没有“总是”播放他们喜欢的音乐的人数的标准差是多少？

70. 一条新的自动生产线平均每天出现故障 1.5 次。由于每次故障都会增加成本，管理部门关心一天发生 3 次或更多次故障的概率。假设故障是随机发生的，任何两段相等时间段内有一次故障发生的概率相等，并且一段时间内是否发生故障与另一段时间是否发生故障是独立的。求一天中有 3 次及以上故障的概率是多少？

72. 顾客们随机而独立地进入一家银行。每分钟内有 1 名顾客来到的概率都相等。假设平均每分钟有 3 名顾客来到，试问：

a. 1 分钟内恰好有 3 名顾客到来的概率是多少？

b. 1 分钟内至少有 3 名顾客到来的概率是多少？

74.《美国新闻与世界报道》对美国的商学院进行排名，哈佛大学和斯坦福大学的商学院并列榜首。此外，排名 7 到 10 名的商学院毕业生的平均成绩点数（GPA）不低于 3.50（“美国最好的商学院”，《美国新闻与世界报道》，2009 年）。假定从排名前十的商学院中随机选取两所。试问：

a. 恰有 1 所商学院学生平均成绩点数（GPA）不低于 3.50 的概率是多少？

b. 两所商学院学生平均成绩点数（GPA）都不低于 3.50 的概率是多少？

c. 两所商学院学生平均成绩点数（GPA）都低于 3.50 的概率是多少？

第6章 连续型概率分布

CHAPTER 6

实践中的统计

宝洁公司[一]

俄亥俄州，辛辛那提市

宝洁公司生产和销售诸如洗涤剂、纸尿布、非处方类的药品、牙膏、肥皂、漱口水和纸巾等产品。在世界各地，它比其他日用消费品公司在各个领域拥有更多的知名品牌。由于收购了吉列公司，宝洁公司还生产和销售剃须刀、刀片以及许多其他个人护理产品。

作为在决策中应用统计方法的领导者，宝洁公司雇用具有各种学术背景的人士：工程学、统计学、运筹学和商务。这些人士提供的定量技术主要有概率决策和风险分析、高级模拟、质量改进和数量方法（如线性规划、回归分析、概率分析）。

宝洁公司的工业化学部是脂肪乙醇的主要供应者，脂肪乙醇从诸如椰子油这样的天然原料和石油衍生物中提取。该部门想知道扩建其脂肪乙醇生产设备的经济风险和机会，于是求助于宝洁公司概率决策和风险分析领域的专家。经过对问题的建模分析，他们确定获利能力的关键在于以石油为原料和以椰子为原料的成本差异。未来成本是未知的，但是分析人员能用下列连续型随机变量对它们进行近似。

x——提取1磅脂肪乙醇所需椰子油的成本；

y——提取1磅脂肪乙醇所需石油原料的成本。

由于赢利能力的关键是这两个随机变量间的差，于是在分析中引入第三个随机变量 $d=x-y$。首先与专家们访谈后确定 x 和 y 的概率分布。然后，根据这些信息建立价格差异 d 的概率分布。这个连续型概率分布表明，价格差不超过0.065 5美元的概率是0.90，价格差不超过0.035美元的概率是0.50。另外，价格差不超过0.004 5美元的概率只有0.10[二]。

工业化学品部认为，将原材料价格差的影响量化是达成共识的关键。利用所得到的概率对原材料的价格差进行灵敏度分析。这种分析将为管理层的决策提供充分的依据。

使用连续型随机变量及其概率分布有助于宝洁公司对其脂肪乙醇产品的经济风险进行分析。本章，你将会学习连续型随机变量及其概率分布的知识，其中包括统计学中最重要的概率分布——正态分布。

上一章中，我们介绍了离散型随机变量及其概率分布。本章，我们转向研究连续型随机变量。特别讨论了三种连续型概率分布：均匀分布、正态分布和指数分布。

离散型随机变量和连续型随机变量之间最根本的区别在于二者在概率计算上是不同的。对一个离散型随机变量，概率函数 $f(x)$ 给出了随机变量 x 取某个特定值的概率。而对连续型随机变量，与概率函数相对应的是**概率密度函数**（probability density function），也记做 $f(x)$。不同的是，概率密度函数并没有直接给出概率。但是，通过给定区间上曲线 $f(x)$ 下的面积给出连续型随机变量在该区间取值的概率。因此，当计算连续型随机变量的概率时，我们计算的是随机变量在某个区间内取值的概率。

由于在任一特定点上曲线 $f(x)$ 下的面积为零，所以连续型随机变量概率的定义意味着它取某一特定值的概率为零。在第6.1节，我们以服从均匀分布的连续型随机变量为例来说明这些概念。

本章的大部分内容致力于描述和展示正态分布的应用。正态分布是非常重要的分布，它具有广泛的适用性并且在统计推断中被广泛使用。最后，介绍了指数分布。在诸如等待时间和服务时间这类因素的应用研究中，常常会用到指数分布。

6.1 均匀概率分布

令随机变量 x 表示某航班从芝加哥飞往纽约的飞行时间。假定飞行时间可以取区间［120，140］中的任意

[一] 作者感谢宝洁公司的Joel Kahn先生，他为“实践中的统计”提供了本案例。

[二] 为保护数据所有权，此处所给出的价格差异是经过改动的。

值。由于随机变量 x 可以在该区间中取任意值，因此 x 不是离散型随机变量而是一个连续型随机变量。假定有足够多的实际飞行数据可以得出结论：对于区间［120，140］内的任意两个 1 分钟长度的子区间，飞行时间在这两个子区间的概率是相同的。[㊀]由于飞行时间在每个 1 分钟长度的子区间内是等可能的，因此称随机变量 x 服从**均匀概率分布**（uniform probability distribution）。飞行时间是服从均匀分布的随机变量，它的概率密度函数为：

$$f(x)=\begin{cases}1/20 & 120\leqslant x\leqslant 140\\ 0 & \text{其他}\end{cases}$$

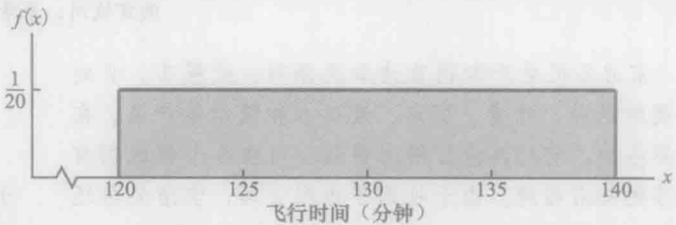


图 6-1 飞行时间的均匀概率分布

图 6-1 是这一概率密度函数的图形。一般情形下，若随机变量 x 服从均匀概率分布，则它的密度函数的公式如下。

均匀概率密度函数

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant x\leqslant b\\ 0, & \text{其他}\end{cases}\tag{6-1}$$

对于飞行时间这一随机变量，$a=120$，$b=140$。

对于一个连续型随机变量，正如本章前言中所提及的那样，我们仅仅考虑随机变量在某个特定区间内取值的概率。在飞行时间的例子中，一个可接受的概率问题是：飞行时间在 120 ~ 130 分钟之间的概率是多大？即 $P(120\leqslant x\leqslant 130)=?$ 由于飞行时间一定介于 120 分钟和 140 分钟之间，并且概率在这一区间上的分布是均匀的，因此我们有理由认为 $P(120\leqslant x\leqslant 130)=0.50$。在下一小节，我们说明这一概率可以通过计算区间［120，130］上曲线 $f(x)$ 下的面积来得到（见图 6-2）。

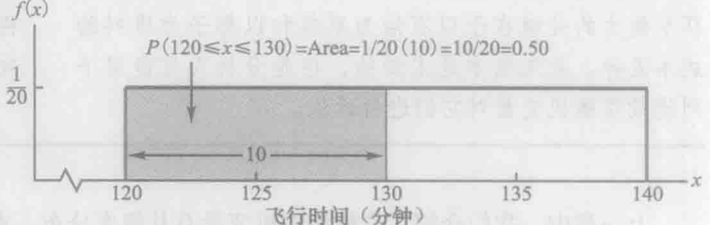


图 6-2 面积是飞行时间在 120 ~ 130 分钟之间的概率

6.1.1 用面积度量概率

观察图 6-2 中区间［120，130］上曲线 $f(x)$ 下的面积。该区域是长方形，长方形的面积等于宽乘高。区间的宽等于 $130-120=10$，高等于概率密度函数的值 $f(x)=1/20$，于是面积 = 宽 × 高 $=10\times 1/20=10/20=0.50$。

对于曲线 $f(x)$ 下的面积和概率，你发现了什么？二者是相同的！事实上，这一结论对所有的连续型随机变量都成立。一旦确定了概率密度函数 $f(x)$，则 x 在区间［x_1，x_2］内取值的概率可通过计算在区间［x_1，x_2］上曲线 $f(x)$ 下的面积得到。

假定飞行时间服从均匀分布并将面积解释为概率，我们就可以回答许多关于飞行时间的概率问题。例如，飞行时间在 128 ~ 36 分钟间的概率是多少？区间宽度是 $136-128=8$，高度 $f(x)=1/20$，于是 $P(128\leqslant x\leqslant 136)=8\times 1/20=0.40$。

注意，$P(120\leqslant x\leqslant 140)=20\times 1/20=1$，即曲线 $f(x)$ 下的总面积等于 1。所有连续型概率分布都具有这一性质，这与离散型概率函数的概率之和必须等于 1 相类似。对连续型概率密度函数，我们还要求对所有的 x 值必须满足 $f(x)\geqslant 0$，这与离散型概率函数要求 $f(x)\geqslant 0$ 是类似的。

处理连续型随机变量和离散型随机变量时，主要存在以下两方面的区别：

（1）我们不再讨论随机变量取某一特定值的概率。取而代之，我们讨论随机变量在某一给定区间上取值的

㊀ 只要概率与区间长度成比例，随机变量就是均匀分布。

概率。

(2) 连续型随机变量在某个给定区间 $[x_1, x_2]$ 上取值的概率被定义为在区间 $[x_1, x_2]$ 上概率密度函数 $f(x)$ 曲线下的面积。因为单点是宽度为零的区间，这意味着连续型随机变量取某个特定值的概率为零[⊖]。这也意味着，无论是否包括端点连续型随机变量在某个区间上取值的概率是相同的。连续型随机变量的数学期望和方差的计算与离散型随机变量类似。然而，由于计算过程涉及积分计算，公式的严格推导参见高级教程。

对于本节介绍的连续型均匀概率分布，数学期望和方差的公式如下

$$E(x) = \frac{a+b}{2} \quad \mathrm{Var}(x) = \frac{(b-a)^2}{12}$$

式中，a 为随机变量所能取的最小值；b 为随机变量所能取的最大值。

航班从芝加哥到纽约的飞行时间服从均匀分布，根据上面两个公式得到：

$$E(x) = \frac{(120+140)}{2} = 130 \quad \mathrm{Var}(x) = \frac{(140-120)^2}{12} = 33.33$$

飞行时间的标准差是方差的算术平方根。于是，$\sigma = 5.77$ 分钟。

注释和评论

为了更清楚地说明为什么概率密度函数的高度不是概率，考虑一个具有下列均匀概率分布的随机变量。

$$f(x) = \begin{cases} 2, & 0 \leqslant x \leqslant 0.5 \\ 0, & 其他 \end{cases}$$

x 在 0 和 0.5 之间取值时，概率密度函数的高度是 2。然而，我们知道概率不可能大于 1。因此，$f(x)$ 并非 x 的概率。

练 习

方法

2. 已知随机变量 x 服从区间 [10, 20] 上的均匀分布。
   a. 绘制概率密度函数图。
   b. 计算 $P(x<15)$。
   c. 计算 $P(12 \leqslant x \leqslant 18)$。
   d. 计算 $E(x)$。
   e. 计算 $\mathrm{Var}(x)$。

应用

4. 大多数计算机语言都有一个能够生成随机数的函数。在 Excel 中，RAND 函数可用于产生 0~1 的随机数。如果令 x 表示用 RAND 函数生成的随机数，那么 x 是具有下列概率密度函数的连续型随机变量。

$$f(x) = \begin{cases} 1, & 0 \leqslant x \leqslant 1 \\ 0, & 其他 \end{cases}$$

   a. 绘制概率密度函数图。
   b. 生成的随机数取值介于 0.25 与 0.75 之间的概率是多少？
   c. 生成的随机数取值不大于 0.30 的概率是多少？
   d. 生成的随机数取值大于 0.60 的概率是多少？
   e. 在 Excel 工作表中的 50 个单元格中输入"=RAND ()"，生成 50 个随机数。
   f. 计算 (e) 中所产生的随机数的均值和标准差。

6. 在 30 分钟的电视情景喜剧中，平均有 22 分钟是节目（CNBC，2006 年 2 月 23 日）。假定节目的时长近似服从 18~26 分钟的均匀概率分布。
   a. 一个情景喜剧播出时，节目时长达到或超过 25 分钟的概率是多少？
   b. 一个情景喜剧播出时，节目时长介于 21~25 分钟的概率是多少？
   c. 一个情景喜剧播出时，商业或其他非节目插播时间超过 10 分钟的概率是多少？

⊖ 为说明随机变量取任一单点值的概率为 0，参见图 6-2。比如，在 $x=125$ 处，$P(x=1.25)=P(125 \leqslant x \leqslant 125)=0 \times 1/20=0$。

6.2 正态概率分布

正态概率分布⊖（normal probability distribution）是描述连续型随机变量的最重要的一种概率分布。正态分布有着广泛的实际应用，比如，人的身高和体重、考试成绩、科学测量、降雨量以及其他类似的数值，都近似服从正态概率分布。正态概率分布也广泛应用于统计推断，而统计推断正是本书其余部分的主要内容。在这些应用中，正态分布描述了抽样可能得到的结果。

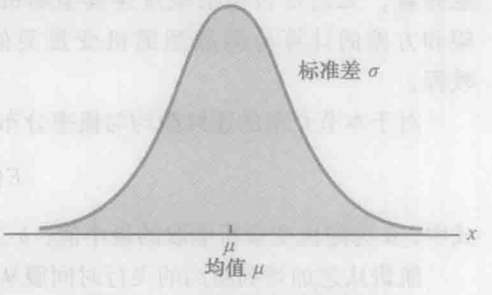


图 6-3 正态分布的钟形曲线

6.2.1 正态曲线

正态分布的形状或形态可以通过图 6-3 中钟形的正态分布曲线来说明。定义正态分布的概率密度函数如下，这是一条钟形曲线。

正态概率密度函数

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2} \tag{6-2}$$

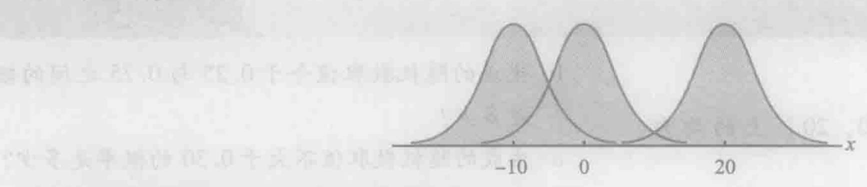
式中，μ 为均值；σ 为标准差；π 为 3.141 59；e 为 2.718 28。

我们观察到，正态分布具有下列特征：

（1）正态分布族中的每个分布因均值 μ 和标准差 σ 这两个参数的不同而不同⊜。

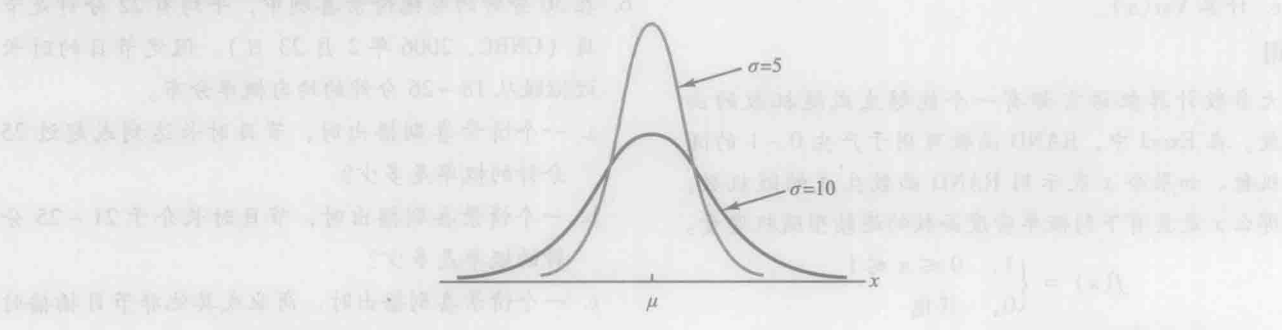
（2）正态曲线的最高点在均值处达到，均值还是分布的中位数和众数。

（3）分布的均值可以是任意数值：负数、零或正数。下图给出了有相同的标准差但有不同均值（－10、0 和 20）的三个正态分布。



（4）正态分布是对称的。均值左边的曲线形状是均值右边的曲线形状的镜像。曲线的尾端向两个方向无限延伸，且理论上永远不会与横轴相交。由于正态分布是对称的，从而它不是偏斜的，偏度为 0。

（5）标准差决定曲线的宽度和平坦程度。标准差越大则曲线越宽、越平坦，表明数据有更大的变异性。下图给出了有相同均值但有不同标准差的两个正态分布。



⊖ 法国数学家棣莫弗（Abraham de Moivre）在 1733 年出版的 *The Doctrine of Chances* 中推导了正态分布。

⊜ 正态曲线的两个参数 μ 和 σ，确定了正态分布的位置和形状。

（6）正态随机变量的概率由正态曲线下的面积给出。正态分布曲线下的总面积是1。由于分布是对称的，曲线下方均值左侧的面积为0.5，曲线下方均值右侧的面积也是0.5。

（7）下面是随机变量在一些常用区间内取值的百分比[1]。

a. 正态随机变量有68.3%的值在均值加减一个标准差的范围内。

b. 正态随机变量有95.4%的值在均值加减两个标准差的范围内。

c. 正态随机变量有99.7%的值在均值加减三个标准差的范围内。

图6-4是对上述性质a、b和c的图示。

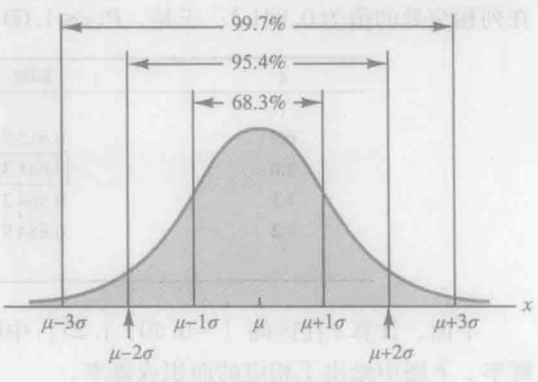


图6-4 正态分布曲线下的面积

6.2.2 标准正态概率分布

如果一个随机变量服从均值为0并且标准差为1的正态分布，则称该随机变量服从**标准正态概率分布**（standard normal probability distribution）。通常用字母z表示这一特殊的正态随机变量。图6-5是标准正态分布的图，它与其他正态分布有相同的形状，不同的是$\mu=0$和$\sigma=1$。对于正态概率密度，曲线的高度是不断变化的，需要采用积分计算曲线下的面积——概率。

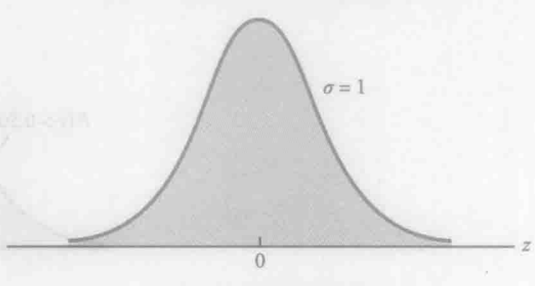


图6-5 标准正态分布

由于$\mu=0$和$\sigma=1$，标准正态概率密度函数的公式如下，它比式（6-2）更为简单。

标准正态密度函数

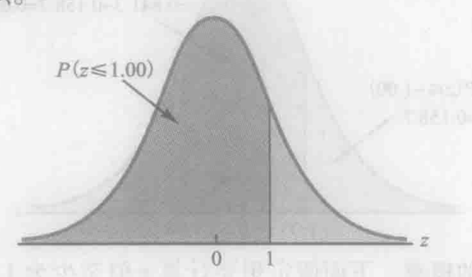
$$f(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$$

与其他连续型随机变量一样，可以通过计算概率密度函数曲线下的面积得出正态分布的概率。于是，为了得到正态随机变量在某个特定区间内的概率，我们必须计算在该区间中正态曲线下的面积。

对于标准正态分布，正态曲线下的面积已计算出来并已编制成可用于计算概率的数学用表。这个表就是附录B中的表1，分别是z值小于或等于均值零的面积（累积概率）；z值大于或等于均值零的面积（累积概率）。

我们需要计算三种类型的概率：（1）标准正态随机变量z小于或等于某个给定值的概率；（2）z在两个给定值之间的概率；（3）z大于或等于某个给定值的概率。通过几个例子，我们说明如何使用标准正态分布累积概率表来计算这三种类型的概率。

首先，说明如何计算标准正态随机变量z小于或等于1.00的概率$P(z\leq1.00)$[2]。这是一个累积概率，等于下图中$z=1.00$左边正态曲线下的面积。



[1] 这些百分比是第3.3节介绍的经验法则的理论依据。

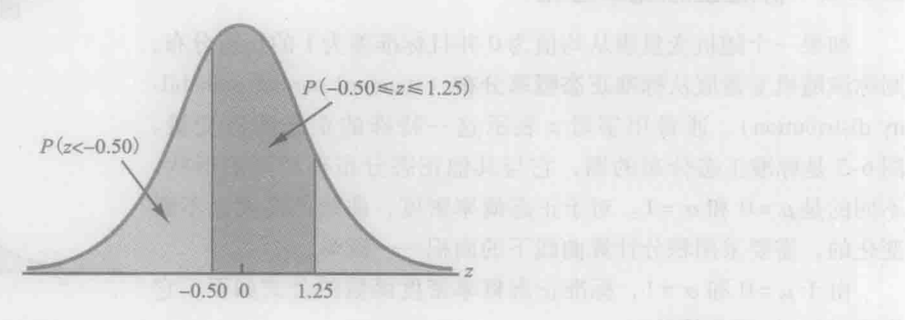
[2] 由于标准正态随机变量是连续的，$P(z\leq1.00)=P(z<1.00)$。

查标准正态概率表，与 $z=1.00$ 对应的累积概率等于表中行标记为 1.0，列标记为 0.00 相交处的数值。我们首先在表左侧第一列中找到 1.0，然后在表的第一行中找到 0.00。通过观察表体，我们发现 1.0 所在行和 0.00 所在列相交处的值为 0.841 3；于是，$P(z\leqslant1.00)=0.841\,3$。下面是用于演示上述步骤的部分概率表。

z	0.00	0.01	0.02
⋮			
0.9	0.815 9	0.818 6	0.821 2
1.0	0.841 3	0.843 8	0.846 1
1.1	0.864 3	0.866 5	0.868 6
1.2	0.884 9	0.886 9	0.888 8
⋮			

$P(z\leqslant1.00)$

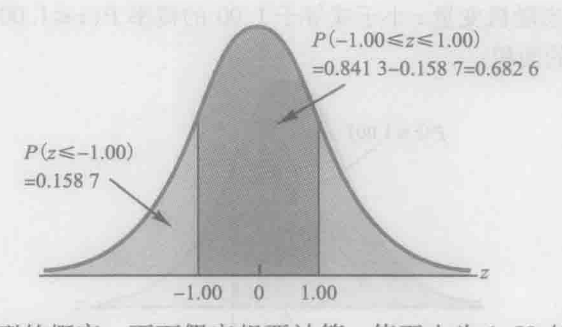
下面，计算 z 在区间 [−0.50，1.25] 中取值的概率 $P(-0.50\leqslant z\leqslant1.25)$。借此演示如何计算第二种类型的概率。下图中给出了相应的面积或概率。



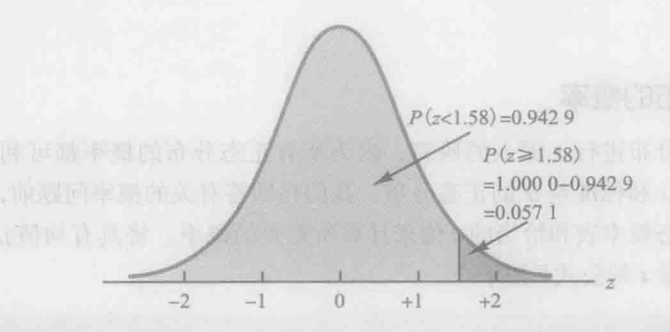
计算这个概率值需要经过三个步骤。首先，计算 $z=1.25$ 左边正态曲线下的面积。然后，计算 $z=-0.50$ 左边正态曲线下的面积。最后，用 $z=1.25$ 左边的面积减去 $z=-0.50$ 左边的面积便得到 $P(-0.50\leqslant z\leqslant1.25)$。

为了得到 $z=1.25$ 左边正态曲线下的面积，我们首先在标准正态概率表中找到 1.2 所在行，然后移到 0.05 所在的列。因为 1.2 所在行和 0.05 所在列相交处的数值为 0.894 4，即 $P(z\leqslant1.25)=0.894\,4$。类似地，查表得 −0.5 所在行和 0.00 所在列相交处的数是 0.308 5，故 $P(z\leqslant-0.50)=0.308\,5$。于是，$P(-0.50\leqslant z\leqslant1.25)=P(z\leqslant1.25)-P(z\leqslant-0.50)=0.894\,4-0.308\,5=0.585\,9$。

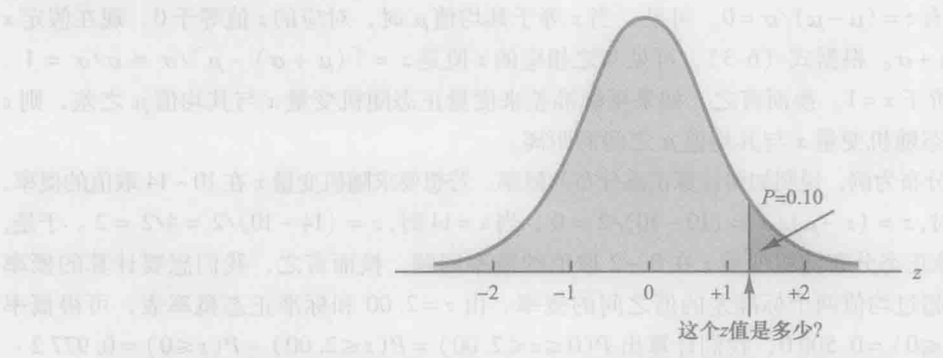
再举另外一个例子，说明如何计算 z 在两个给定值之间取值的概率。人们感兴趣的往往是如何计算正态随机变量的取值在均值一定倍数的标准差范围内的概率。假定，我们想要计算标准正态随机变量在均值附近一个标准差范围内取值的概率，即 $P(-1.00\leqslant z\leqslant1.00)$。为了计算这个概率，我们需要算出在 −1.00 ~ 1.00 曲线下的面积。在前面我们已经得出 $P(z\leqslant1.00)=0.841\,3$。再查表，得到 $z=-1.00$ 左边曲线下的面积为 0.158 7，即 $P(z\leqslant-1.00)=0.158\,7$。于是，$P(-1.00\leqslant z\leqslant1.00)=P(z\leqslant1.00)-P(z\leqslant-1.00)=0.841\,3-0.158\,7=0.682\,6$。下图是这个概率的图示。



为了说明如何计算第三种类型的概率，下面假定想要计算 z 值至少为 1.58 的概率，即 $P(z\geqslant1.58)$。累积正态表中 $z=1.5$ 所在行和 0.08 所在列交叉处的值为 0.942 9，于是 $P(z<1.58)=0.942\,9$。然而，由于正态曲线下总面积等于 1，从而 $P(z\geqslant1.58)=1-0.942\,9=0.057\,1$。下图是对这种情况的一个图示。



在前面的例子中，我们演示了如何对于给定的 z 值计算概率。在某些情况下，给定一个概率，我们感兴趣的问题是反过来求相应的 z 值。假定，我们想要求某个 z 值，使得大于 z 值的概率为0.10。下图是对这种情况的一个图示。



这一问题与前面的例子正好相反。前面，我们设定了感兴趣的 z 值，然后找相应的概率或面积。本例中，我们给定概率或面积，要求找到相应的 z 值。为此，我们以不同的方式使用标准正态概率表[㊀]。

标准正态概率表给出了一个特定 z 值左边正态曲线下的面积。我们已经获得的信息是，曲线上侧尾端的面积是0.10。因此，在未知 z 值左边正态曲线下的面积是0.900 0。查概率表发现，0.899 7是最接近0.900 0的累积概率值。用于得出这一结果的部分概率表如下。

z	0.06	0.07	0.08	0.09
⋮				
1.0	0.855 4	0.857 7	0.859 9	0.862 1
1.1	0.877 0	0.879 0	0.881 0	0.883 0
1.2	0.896 2	0.898 0	0.899 7	0.901 5
1.3	0.913 1	0.914 7	0.916 2	0.917 7
1.4	0.927 9	0.929 2	0.930 6	0.931 9
⋮				

表中最接近0.900 0的概率

从表最左边一列和最上一行读 z 值，我们发现相应的 z 值是1.28。这样，$z=1.28$ 左边的面积近似为0.900 0（实际为0.899 7）[㊁]。根据问题最初的提法，z 值大于1.28的概率大约为0.10。

这些例子说明，标准正态概率分布的累积概率表可用于求出有关标准正态随机变量 z 的概率。可以回答两类问题。第一类问题是，指定 z 的一个或多个值，要求我们利用表确定相应的面积或概率。第二类问题是，给定一个面积或概率，要求我们利用概率表确定相应的 z 值。于是，我们需要灵活地运用标准正态概率表来回答感兴趣的概率问题。在大多数情况下，绘制标准正态概率分布图并用恰当的阴影表示相应的面积，将有助于使问题形象

㊀ 给定某个已知的概率值，可以利用标准正态分布概率表反查出相应的 z 值。

㊁ 我们可以在表体中使用插值法得到与面积0.900 0相对应的最接近的 z 值。这样做可以给出更精确（有更多的小数位）的 z 值，$z=1.282$。然而，在大多数实际情况中，通过简单地用最接近概率的表值便可达到足够的精确度。

化，从而帮助得出正确的答案。

6.2.3 计算正态概率分布的概率

上一节中我们对标准正态分布进行了深入的研究，因为所有正态分布的概率都可利用标准正态分布来计算。也就是说，对于具有任意均值μ和标准差σ的正态分布，我们在回答有关的概率问题前，首先应将其转换成标准正态分布，然后再使用标准正态概率表和恰当的z值来计算所需要的概率。将具有均值μ和标准差σ的正态随机变量x转换为标准正态随机变量z的公式[⊖]如下。

转换为标准正态随机变量

$$z=\frac{x-\mu}{\sigma} \tag{6-3}$$

当x的值等于其均值μ时，有$z=(\mu-\mu)/\sigma=0$。可见，当x等于其均值μ时，对应的z值等于0。现在假定x大于其均值一个标准差，即$x=\mu+\sigma$。根据式（6-3），可见与之相应的z值是$z=[(\mu+\sigma)-\mu]/\sigma=\sigma/\sigma=1$。因此，$x$比均值大一个标准差等价于$z=1$。换而言之，如果用标准差来度量正态随机变量$x$与其均值$\mu$之差，则$z$是以$x$的标准差为度量单位的正态随机变量$x$与其均值$\mu$之间的距离。

下面以$\mu=10$和$\sigma=2$的正态分布为例，说明如何计算正态分布的概率。若想要求随机变量x在10～14取值的概率，则根据式（6-3）可得：当$x=10$时，$z=(x-\mu)/\sigma=(10-10)/2=0$；当$x=14$时，$z=(14-10)/2=4/2=2$。于是，$x$在10～14取值的概率等价于标准正态分布随机变量$z$在0～2取值的概率问题。换而言之，我们想要计算的概率是随机变量x取值位于其均值和超过均值两个标准差的值之间的概率。由$z=2.00$和标准正态概率表，可得概率$P(z\leqslant2.00)=0.977\,2$。由于$P(z\leqslant0)=0.500\,0$，我们计算出$P(0\leqslant z\leqslant2.00)=P(z\leqslant2.00)-P(z\leqslant0)=0.977\,2-0.500\,0=0.477\,2$。因此，$x$在10～14取值的概率是0.477 2。

6.2.4 Grear 公司轮胎的问题

现在我们转向研究正态概率分布的应用。假定 Grear 公司刚刚开发了一种新的钢带子午线轮胎，将通过全美连锁折扣商店出售。由于这种轮胎是一种新产品，Grear 公司经理认为，提供轮胎可行驶里程的质保是影响产品被接受的一个重要因素。在最终落实轮胎行驶里程保证政策之前，Grear 公司经理需要有关轮胎的行驶里程x的概率信息。

根据对轮胎的实际道路测试，Grear 公司的工程小组估计轮胎可行驶里程的均值$\mu=36\,500$英里，标准差$\sigma=5\,000$。另外，收集的数据表明正态分布是一个合理的假设。行驶里程超过40 000英里的轮胎所占的百分比有多大？换句话说，轮胎行驶里程x超过40 000英里的概率是多少？这一问题可以通过计算图6-6中深色阴影区域的面积来回答。

当$x=40\,000$时，有：

$$z=\frac{x-\mu}{\sigma}=\frac{40\,000-36\,500}{5\,000}=\frac{3\,500}{5\,000}=0.70$$

现在，观察图6-6的底部，可见与 Grear 轮胎公司正态分布值$x=40\,000$相对应的标准正态分布值$z=0.70$。利用标准正态概率表，在$z=0.70$左边正态曲线下的面积为0.758 0。因此，z大于0.70的概率为$1-0.758\,0=0.242\,0$，从而x超过40 000的概率为

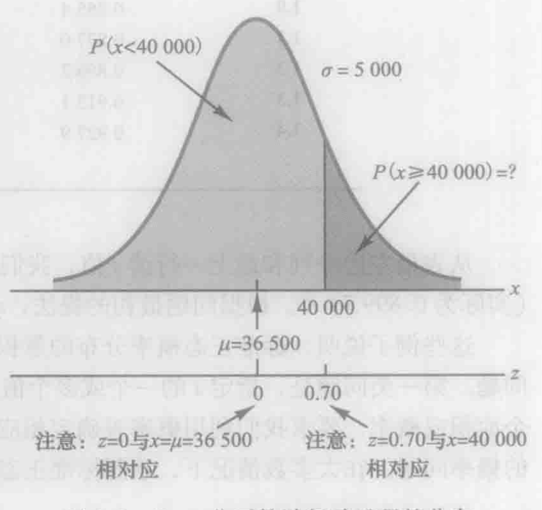


图 6-6 Grear 公司轮胎行驶里程的分布

⊖ 标准正态随机变量的公式类似于我们在第3章介绍的计算数据集的z得分的公式。

0.2420。我们可以得出结论，有大约24.2%的轮胎其行驶里程将超过40000英里。

下面，假设Grear公司正在考虑实施一项质量保证：如果轮胎的行驶里程没有达到质量保证规定的里程，公司将以折扣价提供更换轮胎的服务。如果Grear公司希望达到折扣质量保证条件的轮胎不要超过总数的10%，则应设定的质保里程为多少英里？图6-7是关于该问题的一个图示。

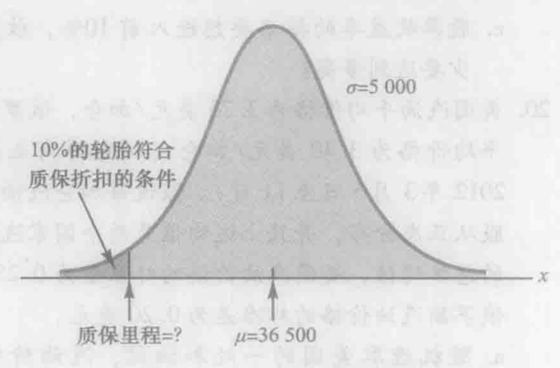


图6-7 Grear公司的质保折扣

根据图6-7，在质保里程值（未知）左边正态曲线下的面积必须为0.10。所以，我们必须找到一个z值使得在它左边标准正态曲线下的面积为0.10。利用标准正态概率表，我们看到$z=-1.28$时，下尾的面积为0.10。因此，$z=-1.28$是与Grear轮胎正态分布所希望的质保里程相对应的标准正态随机变量的值。为了求出与$z=-1.28$相对应的x值，我们有

$$z=\frac{x-\mu}{\sigma}=-1.28$$

$$x-\mu=-1.28\sigma$$

$$x=\mu-1.28\sigma^{㊀}$$

因为$\mu=36\,500$和$\sigma=5\,000$，于是得

$$x=36\,500-1.28\times5\,000=30\,100$$

因此，若设定质量保证为30100英里，则大约有10%的轮胎符合质量保证条件的要求㊁。根据这一信息，也许公司会把轮胎质保里程设定为30000英里。

我们再次看到了概率分布在提供决策所需信息方面所起的重要作用。即一旦对某一特定应用问题确定了其概率分布，就可以取得有关问题的概率信息。概率并不直接给出决策建议，但它能提供帮助决策者更好地了解与问题相关的风险和不确定性的信息。最终，这些信息可以辅助决策者作出正确的决策。

练 习

方法

8. 利用图6-4，绘制均值$\mu=100$和标准差$\sigma=10$的随机变量x的正态曲线。在横轴上标明值70，80，90，100，110，120和130。

10. 绘制标准正态分布图，并且在横轴上标明数值-3，-2，-1，0，1，2和3，然后利用附录B中标准正态分布概率表计算下列概率。
    a. $P(z\leqslant1.5)$ b. $P(z\leqslant1)$
    c. $P(1\leqslant z\leqslant1.5)$ d. $P(0<z<2.5)$

12. 已知z是一个标准正态随机变量，计算下列概率。
    a. $P(0\leqslant z\leqslant0.83)$ b. $P(-1.57\leqslant z\leqslant0)$
    c. $P(z>0.44)$ d. $P(z\geqslant0.23)$
    e. $P(z<1.20)$ f. $P(z\leqslant-0.71)$

14. 已知z是一个标准正态随机变量，对下面各种情况求z值。
    a. z左侧的面积是0.9750。
    b. 0和z间的面积是0.4750。
    c. z左侧的面积是0.7291。
    d. z右侧的面积是0.1314。
    e. z左侧的面积是0.6700。
    f. z右侧的面积是0.3300。

16. 已知z是一个标准正态随机变量，对下面各种情况求z值。
    a. z右侧的面积是0.01。
    b. z右侧的面积是0.025。
    c. z右侧的面积是0.05。
    d. z右侧的面积是0.10。

应用

☆18. 在2009～2011年这3年中，国内大盘股基金的平均收益率为14.4%（《美国个人投资者协会日报》，2012年2月）。假定基金三年的收益率服从正态分布，标准差是4.4%。

㊀ 可见，质量保证里程x位于均值下方距离均值1.28个标准差的位置。
㊁ 符合质保条件的实际百分比为9.68%。

a. 一只大盘国内股票基金，它三年中的收益率至少达到20%的概率有多大？

b. 一只大盘国内股票基金，它三年中的收益率不高于10%的概率有多大？

c. 股票收益率的排名要想进入前10%，收益率至少要达到多高？

20. 美国汽油平均价格为3.73美元/加仑，俄罗斯汽油平均价格为3.40美元/加仑（《彭博商业周刊》，2012年3月5日至11日）。假设每加仑汽油的价格服从正态分布，并且上述均值是两个国家汽油价格的总体均值，美国汽油价格的标准差为0.25美元，俄罗斯汽油价格的标准差为0.20美元。

a. 随机选取美国的一处加油站，汽油价格低于3.50美元/加仑的概率是多少？

b. 俄罗斯的加油站中，汽油价格低于3.50美元/加仑的加油站占多大的百分比？

c. 随机选取俄罗斯的一处加油站，求其汽油价格高于美国汽油价格的均值的概率。

22. 据尼尔森公司报道，家庭每天用于收看电视节目的平均时间再创新高，达到8.35个小时(《今日美国》，2009年11月11日)。假设每个家庭收看电视节目的时间服从正态概率分布，标准差为2.5小时。试问：

a. 一个家庭用于收看电视节目的时间介于5~10个小时的概率是多少？

b. 若一个家庭每天收看电视节目的时间排名处于前3%的位置，则每天至少收看多少个小时？

c. 一个家庭每天收看电视节目的时间多于3个小时的概率是多少？

24. 在纽约证券交易所的每个交易日中，最初的半个小时（清晨）和最后的半个小时（午后）的交易量最大。下面是1月和2月共13天中，清晨的交易量（单位：100万股）数据（《巴伦周刊》，2006年1月23日、2006年2月13日和2006年2月27日)。

214	163	265	194	180
202	198	212	201	
174	171	211	211	

交易量近似服从正态概率分布。

a. 计算均值和标准差，用来估计总体的均值和标准差。

b. 随机选取一天，清晨交易量少于1.8亿股的概率有多大？

c. 随机选取一天，清晨交易量超过2.3亿股的概率有多大？

d. 若某天清晨的交易量跻身最繁忙的前5%交易日，求交易量是多少？

6.3 二项概率的正态近似

在第5.5节我们介绍了离散型的二项概率分布。二项试验由一系列 n 个相同的独立试验组成，每个试验有两种可能结果：成功或失败。所有的试验中成功的概率都是相同的，用 p 来表示。二项分布中随机变量是 n 次试验中成功的次数，在 n 次试验中有 x 次成功的概率是我们关心的概率问题。

当试验次数很大时，用手工计算或计算器求解二项概率函数都是很困难的。在 $np \geq 5$ 和 $n(1-p) \geq 5$ 情况下，正态分布是对二项概率分布的一个简便易行的近似。当使用正态分布近似二项分布时，正态曲线中取 $\mu = np$ 和 $\sigma = \sqrt{np(1-p)}$。

下面，我们举例说明二项分布的正态近似。假定历史经验表明，某公司发票出错的概率为10%。现选取100张发票组成一个样本，我们想计算恰好有12张发票有错的概率，即想计算100次试验中恰好有12次成功的二项概率。在应用二项分布的正态近似时，令 $\mu = np = 100 \times 0.1 = 10$ 和 $\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.1 \times 0.9} = 3$。$\mu = 10$ 和 $\sigma = 3$ 的正态分布如图6-8所示。

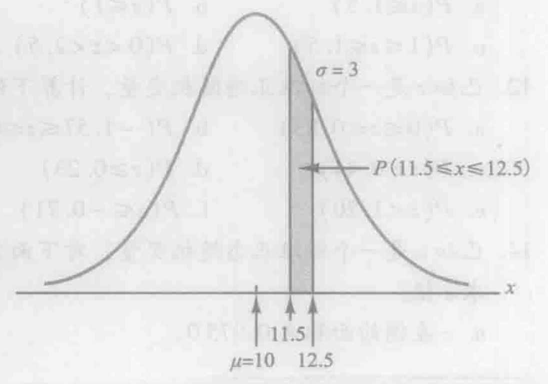


图6-8 一个二项概率分布的正态近似（$n = 100$ 和 $p = 0.10$，有12次出错的概率）

对连续型概率分布，概率是通过计算概率密度函数下方的面积得出的。因此，随机变量取任意单个值的概率为0。为了对恰好有12次成功的二项概率进行近似，我们必须计算11.5和12.5之间正态曲线下的面积。其中11.5和12.5是将12加减0.5得到的，我们称0.5为**连续性校正因子**（continuity correction factor）。这是由于我

们在用一个连续分布来近似一个离散分布。从而，离散型二项分布 $P(x=12)$ 可用连续型正态分布 $P(11.5\leqslant x\leqslant 12.5)$ 来近似。

转换成标准正态分布后计算 $P(11.5\leqslant x\leqslant 12.5)$，有

$$当\ x = 12.5\ 时, z = \frac{x-\mu}{\sigma} = \frac{12.5-10.0}{3} = 0.83$$

和

$$当\ x = 11.5\ 时, z = \frac{x-\mu}{\sigma} = \frac{11.5-10.0}{3} = 0.50$$

利用标准正态概率表，可得曲线下方12.5左侧区域的面积为0.796 7（见图6-8）。类似地，曲线下方11.5左侧区域的面积为0.691 5。这样，11.5与12.5之间的面积是0.796 7 − 0.691 5 = 0.105 2。利用正态近似可得，100次试验中恰好有12次成功的概率等于0.105 2。

再比如，在由100张发票组成的样本中，假定我们想计算有不多于13张发票出错的概率。这一概率近似等于图6-9中正态曲线下的面积。注意：利用连续性校正因子，根据值13.5计算所求的概率。与 $x=13.5$ 相对应的 z 值是

$$z = \frac{13.5-10.0}{3.0} = 1.17$$

标准正态概率表显示，标准正态曲线下 $z=1.17$ 左侧区域的面积是0.879 0。出错次数不多于13次的概率约等于正态曲线下（图6-9中阴影部分）的面积。

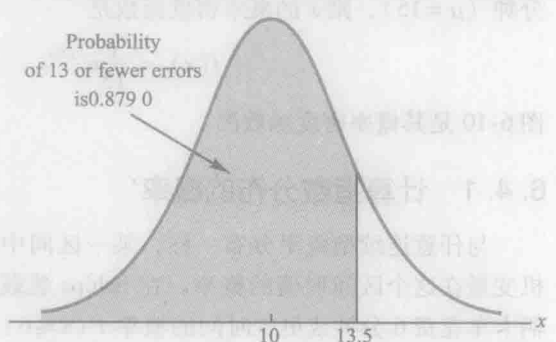


图6-9　二项概率分布的正态近似（$n=100$ 和 $p=0.10$，出错次数不多于13次的概率）

练　习

方法

☆26. 一个二项概率分布，其中 $p=0.20$，$n=100$。试问：

a. 均值和标准差各是多少？

b. 这种情况下，二项概率可以用正态概率分布来近似吗？为什么？

c. 恰好有24次成功的概率是多少？

d. 成功次数在18~22次的概率是多少？

e. 成功次数不多于15次的概率是多少？

应用

☆28. 虽然研究持续表明，吸烟会导致严重的健康问题，但是美国仍有20%的成年人吸烟。在由250名成年人组成的小组中。

a. 成年人中吸烟人数的期望值是多少？

b. 吸烟人数少于40人的概率是多少？

c. 吸烟人数在55~60人的概率是多少？

d. 吸烟人数不少于70人的概率是多少？

30. 影碟和电脑游戏非常盛行，超过70%的家庭玩这些游戏。在玩影碟和电脑游戏的人中，年龄在18岁及以下的人占18%；年龄在18~59岁的人占53%；年龄在59岁以上的人占29%（《华尔街日报》，2012年3月6日）。

a. 由800名玩游戏的人组成一个样本，你预期多少人的年龄在18岁以下？

b. 由600名玩游戏的人组成一个样本，求年龄在18岁以下的人数少于100人的概率。

c. 由800名玩游戏的人组成一个样本，求年龄在18~59岁的人数达到或超过200人的概率。

6.4　指数概率分布

指数概率分布（exponential probability distribution probability distribution）可用于描述诸如到达某洗车处的两辆车的时间间隔、装载一辆卡车所需时间、高速公路上两起重大事故发生地之间的距离等随机变量。[⊖]指数概率密度

⊖　在排队论中，指数分布常用于描述服务时间。

函数如下：

指数概率密度函数

$$f(x) = \frac{1}{\mu}e^{-x/\mu}, x \geqslant 0 \tag{6-4}$$

式中，μ 为期望值或均值。

作为指数概率分布的一个例子，假定在 Schips 码头装载一辆卡车所需时间 x 服从指数分布。如果装车时间的均值或平均所需时间是 15 分钟（$\mu=15$），则 x 的概率密度函数是

$$f(x) = \frac{1}{15}e^{-x/15}$$

图 6-10 是其概率密度函数图。

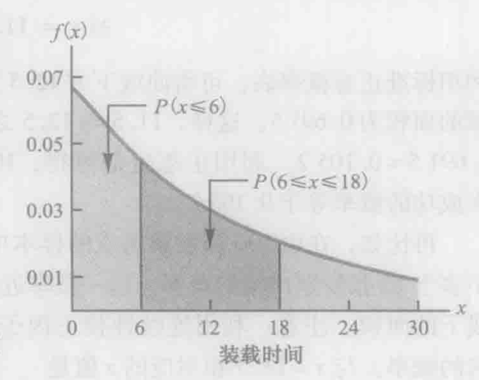


图 6-10 SCHIPS 装载码头例子的指数分布

6.4.1 计算指数分布的概率

与任意连续型概率分布一样，某一区间中相应曲线下的面积是随机变量在这个区间取值的概率。在 Schips 装载码头的例子中，装载一辆卡车花费 6 分钟或更少时间的概率 $P(x\leqslant6)$ 等于图 6-10 中从 $x=0$ 到 $x=6$ 的曲线下区域的面积。类似地，装载一辆卡车花费 18 分钟或更少时间的概率 $P(x\leqslant18)$ 等于从 $x=0$ 到 $x=18$ 的曲线下区域的面积。注意，装载一辆卡车花费在 6 分钟到 18 分钟之间的概率 $P(6\leqslant x\leqslant18)$ 等于从 $x=6$ 到 $x=18$ 的曲线下的面积。

为了计算诸如刚才描述的这些指数分布，我们使用下列公式。它给出指数随机变量取小于或等于某一特定值 x_0 的累积概率。

指数分布：累积概率

$$P(x \leqslant x_0) = 1 - e^{-x_0/\mu} \tag{6-5}$$

在 Schips 装载码头的例子中，x = 装载时间（以分钟计），$\mu=15$ 分钟，根据式（6-5）有

$$P(x \leqslant x_0) = 1 - e^{-x_0/15}$$

因此，装载一辆卡车所用时间不多于 6 分钟的概率是

$$P(x \leqslant 6) = 1 - e^{-6/15} = 0.3297$$

利用式（6-5），我们计算出装载一辆卡车所用时间不多于 18 分钟的概率是

$$P(x \leqslant 18) = 1 - e^{-18/15} = 0.6988$$

因此，装载一辆卡车所用时间在 6～18 分钟的概率等于 0.698 8 − 0.329 7 = 0.369 1。类似地，可以计算所用时间在任意其他区间的概率。

在上面的例子中，装载一辆卡车平均需要的时间为 $\mu=15$ 分钟。指数分布的性质之一是，分布的均值和标准差相等。因此，装载一辆卡车所需时间的标准差为 $\sigma=15$ 分钟，方差为 $\sigma^2=15^2=225$。

6.4.2 泊松分布与指数分布的关系

在第 5.6 节我们介绍了离散概率分布——泊松分布。在研究一个事件在一特定时间段或空间中发生次数时，泊松分布非常有用。泊松分布的概率函数如下：

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

式中，μ 为在某一特定区间发生次数的数学期望或均值。连续型指数概率分布与离散型泊松分布是相互联系的，

泊松分布描述了每一区间中事件发生的次数，指数分布描述了事件发生的时间间隔长度。[⊖]

为了举例说明这种关系，假定在一小时当中到达某一洗车处的汽车数可以用泊松概率分布描述，其均值为每小时10辆。泊松概率函数给出了每小时有x辆汽车到达的概率

$$f(x)=\frac{10^{x}e^{-10}}{x!}$$

由于车辆到达的平均数是每小时10辆，则两车到达的时间间隔的均值为

$$\frac{1\text{ 小时}}{10\text{ 辆车}}=0.1\text{ 小时 / 辆}$$

于是，描述两车到达的时间间隔的对应分布是指数分布，其均值$\mu=0.1$小时/辆，从而指数概率密度函数为

$$f(x)=\frac{1}{0.1}e^{-x/0.1}=10e^{-10x}$$

注释和评论

正如我们在图6-10中所看到的那样，指数分布是右偏的。确实，指数分布的偏度为2。指数分布让我们更直观地看到了偏态分布的特点。

练习

方法

32. 考虑下列指数概率密度函数

$$f(x)=\frac{1}{8}e^{-x/8},\quad x\geqslant 0$$

a. 计算$P(x\leqslant 6)$。 b. 计算$P(x\leqslant 4)$。

c. 计算$P(x\geqslant 6)$。 d. 计算$P(4\leqslant x\leqslant 6)$。

应用

34. Droid Razr Maxx的摩托罗拉的一款手机，当主要用于通话时充电电池可使用20个小时（《华尔街日报》，2012年3月7日），当主要通过蜂窝数据上网时充电电池使用时间下降到7个小时。假定上述两种情形下，电池的使用时间都服从指数分布。

a. 当主要用于通话时，求Droid Razr Maxx手机电池使用时间的概率密度函数。

b. 随机选取一部Droid Razr Maxx手机，当主要用于通话时求电池使用时间不长于15个小时的概率。

c. 随机选取一部Droid Razr Maxx手机，当主要用于通话时求电池使用时间超过20个小时的概率。

d. 随机选取一部Droid Razr Maxx手机，当主要用于上网时求电池使用时间不长于5个小时的概率。

36. Comcast是美国最大的有线电视企业，第二大互联网服务供应商、第四大电信服务供应商。它虽然以优质而可靠的服务闻名，但是仍可能定期会发生不可预期的服务中断。2009年1月14日，佛罗里达州南部的Comcast用户遭遇了服务中断。当用户给Comcast办公室打电话时，留言信息告诉他们公司已经知道发生了服务中断故障，预计2个小时内能恢复服务。假设平均修理时间为2个小时，并且修理时间服从指数概率分布。求：

a. 有线服务在1个小时之内修好的概率是多少？

b. 维修历时1~2个小时之间的概率是多少？

c. 一名用户在下午1点钟给Comcast的办公室打电话，用户的有线服务直至下午5点仍未恢复的概率是多少？

38. 工作时被干扰是否会降低你的工作效率？根据加州大学欧文分校的研究，商务人士被干扰的频率大约是每小时5.5次（《财富》，2006年3月20日）。假设干扰次数服从泊松分布。

a. 求两次干扰之间的时间间隔的概率分布。

b. 在15分钟内，某商务人士不被干扰的概率是多少？

c. 在10分钟内，某商务人士被干扰的概率是多少？

⊖ 如果到达的数目服从泊松分布，那么两次到达之间的等待时间服从指数分布。

小结

本章把对概率分布的讨论推广到连续型随机变量的情形。离散型和连续型概率分布在概念上的主要区别在于计算概率的方法不同。对离散型分布，概率函数$f(x)$给出随机变量x取各种值的概率。对连续型分布，概率密度函数$f(x)$没有直接给出概率值，概率密度函数$f(x)$图或者曲线下的面积才是概率。由于曲线下单个点的面积是0，所以连续型随机变量取某一特定值的概率是0。

我们详细介绍了三种连续型概率分布——均匀分布、正态分布和指数分布。其中，正态分布广泛应用于统计推断，并且在本书的其余内容中将有广泛的应用。

关键术语

probability density function 概率密度函数 用于计算连续型随机变量概率的函数。用某一区间中概率密度函数曲线下的面积来表示概率。

uniform probability distribution 均匀概率分布 一种连续型概率分布，随机变量在等长度的区间上取值的概率相同。

normal probability distribution 正态概率分布 一种连续型概率分布，其概率密度函数呈钟形，由均值μ和标准差σ确定。

standard normal probability distribution 标准正态概率分布 均值为0并且标准差为1的正态分布。

continuity correction factor 连续性修正因子 当用连续型正态概率分布来近似离散型二项概率分布时，从x中加上或减去的值0.5。

exponential probability distribution 指数概率分布 一种连续型概率分布，用于计算完成一项任务所需时间的概率。

重要公式

均匀概率密度函数

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant x\leqslant b\\ 0, & 其他\end{cases} \tag{6-1}$$

正态概率密度函数

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2} \tag{6-2}$$

标准正态随机变量的变换

$$z=\frac{x-\mu}{\sigma} \tag{6-3}$$

指数概率密度函数

$$f(x)=\frac{1}{\mu}e^{-x/\mu},\ x\geqslant 0 \tag{6-4}$$

指数分布：累积概率

$$P(x\leqslant x_0)=1-e^{-x_0/\mu} \tag{6-5}$$

补充练习

40. 据美国大学生体育协会估计，正式公办大学的全额体育奖学金为每年19 000美元（《华尔街日报》，2012年3月12日）。假设奖学金服从标准差为2 100美元的正态分布。
    a. 若奖学金金额的排名位于后10%，求奖学金的金额是多少？
    b. 整个奖学金中。数额不低于22 000美元的奖学金占多大百分比？
    c. 若获得排名前3%的奖学金，求奖学金的金额是多少？

42. 2012年年初，经济困难开始对法国福利体系的额度产生影响。每天去巴黎一家典当行抵押物品的人数增加到658人，典当行抵押物品的人数可以作为衡量困难程度的指标（《彭博商业周刊》，2012年3月5日~3月11日）。假定2012年每天去典当行抵押物品的人数服从均值为658的正态分布。
    a. 假定在3%的日子里，抵押物品的人数不多于610。求每天去典当行抵押物品人数的标准差。
    b. 某天当中，去典当行抵押物品人数在600~700人的概率是多少？
    c. 在典当行最为繁忙的3%的日子里，去典当行抵

押物品人数达到多少？

44. Ward Doering 自动售货公司正考虑提供一项特殊服务合同，以负担服务工作所要求的设备租赁成本。根据经验，公司经理估计年服务成本近似服从正态分布，其均值为150美元，标准差为25美元。
    a. 如果公司以每年200美元的价格向客户提供这种服务，则一名客户服务成本超过合同价格200美元的概率是多少？
    b. Ward公司服务合同的期望利润是多少？

46. 假设某大学入学考试的考试成绩服从正态分布，其均值为450，标准差为100。
    a. 考试分数在400~500的人数占多大百分比？
    b. 假定某人得分630，比此人考试分数高的考生的百分比有多大？比此人考试分数低的考生的百分比有多大？
    c. 如果某大学不招收分数在480分以下的学生，参加考试的学生中被该大学接受的百分比是多少？

48. 一种机器向容器填充某种产品。根据过去的数据已知填充量的标准差是0.6盎司。如果容器中只有2%的容量低于18盎司，这种机器填充量的均值是多少？即应该等于多少？假设填充量服从正态分布。

50. 一名在拉斯维加斯赌场的黑杰克玩家得知，如果他以平均每手50美元赌资玩4小时的话，则店里提供免费房间。他的赌博策略使得在每一手中有0.49的概率取胜。玩家知道每小时大约赌60手。假定他以每手50美元赌资玩4小时。
    a. 玩家的期望支付额是多少？
    b. 玩家输掉1 000美元或更多的概率是多少？
    c. 玩家赢的概率是多少？
    d. 假设玩家有1 500美元，则他输掉所有的钱概率是多少？

52. 平均每分钟大约有7个人登录北美洲含早餐服务旅馆的网站（Time，2001年9月）。假定每分钟该网站的浏览人数服从泊松概率分布。
    a. 该网站相邻两次访问之间的平均时间间隔是多少？
    b. 给出该网站相邻两次访问之间的时间间隔所服从的指数概率密度函数。
    c. 在1分钟内无人访问该网站的概率是多少？
    d. 在12秒内无人访问该网站的概率是多少？

54. 某保险索赔办公室称，两次电话呼叫的时间间隔（以分钟计）服从如下指数概率分布：

$$f(x) = 0.50e^{-0.50x}, \quad x \geqslant 0$$

    a. 两次电话之间的平均等待时间是多少？
    b. 两次电话之间的时间间隔不多于30秒的概率是多少？
    c. 两次电话之间的时间间隔不多于1分钟的概率是多少？
    d. 有5分钟或更长时间没有电话呼叫的概率是多少？

案例 Specialty玩具公司

Specialty玩具公司销售大量新款又有创意的玩具。经营者知道节日前是推出新款玩具的最佳时机，因为许多家庭此时都要为12月里的节日准备礼物。Specialty公司发现一种新款玩具颇具市场潜力，于是决定从10月份起将其投放市场。

为保证玩具能在10月份在商场按期到货，Specialty公司在每年的6月或者7月向制造商下达订单。儿童玩具的需求量瞬息万变。如果一种新款玩具推出时正值市面上缺货，那么可能会有很高的销售量，从而得到大笔的利润。但是，推出一种新款玩具时也可能会遇到滞销，这将使得Specialty公司积压大量的存货，从而不得不降价销售。因此，公司所面临的最主要的问题是：为保证节日期间的供应，应下达多大数目的订单？如果订货量太少，可能丧失销售收入；反之如果订货量过大，则可能由于低价清货而降低利润。

Specialty公司计划下一季度向市场推出一款名为天气熊的新产品。这种会说话的玩具熊由中国台湾地区一家公司生产。当孩子握小熊的手的时候，玩具熊中内置的气压计作出反应，小熊便开口说话预测天气情况。小熊共有“天气真不错，好好玩吧”以及“我想今天也许有雨，别忘了带伞”等五种说法。产品检验发现，虽然这不是一个专门的天气预报装置，但它的预测还是相当不错的。Specialty公司的一些管理者甚至认为，玩具熊对天气的预报可以和当地电视台的天气预报媲美。

与其他产品一样，Specialty公司必须确定为即将到来的下一季中的节日应下达多少只天气熊的订单。管理层

团队建议准备15 000只、18 000只、24 000只或者28 000只。不同的订货方案说明管理层团队中对产品的市场潜力存在相当大的分歧。生产管理部门要求你分析不同订货方案的存货出清概率，估计潜在利润并推荐一种订货方案。每个天气熊的成本为16美元，Specialty公司希望以24美元的价格出售。如果节日过后仍有存货，Specialty公司将以每只5美元的价格清仓销售。根据以往同类产品的历史销售量，Specialty公司产品销量的资深预测专家预计天气熊的需求量为20 000只，需求量在10 000只到30 000只之间的概率为0.95。

管理报告

准备一份管理报告说明下列问题，并针对天气熊产品推荐你的一种订货方案。

1. 根据销量预测专家的预测，需求量的分布近似服从正态概率分布。简略说明该分布，给出其均值和标准差。

2. 计算管理团队所推荐的订货方案的存货出清概率。

3. 在下列三种情形的销售量下，分别计算管理团队所推荐的订货方案的预计利润。最坏的情形：销售量为10 000只；最可能出现的情形：销售量为20 000只；最好的情形：销售量为30 000只。

4. Specialty公司的一名经理认为产品的潜在利润很高，因此订货应该以70%的概率保证满足市场需求，仅需以30%的概率出现脱销。在这种情形下，天气熊的订货量应该是多少？在上述三种不同的销售量下，预计利润分别是多少？

5. 给出你的订货方案及其预计利润，并对订货方案的合理性作出解释。

第7章 抽样和抽样分布

CHAPTER 7

实践中的统计

MeadWestvaco 有限公司[⊖]
康涅狄格州，斯坦福德

MeadWestvaco 公司是一家生产包装纸、铜版纸和特种纸以及特种化学制品的企业，是同行业中的领军人物。公司雇用的员工超过1.7万名，在全世界30个国家有业务活动，服务的客户遍布大约100个国家。MeadWestvaco 内部的顾问组通过抽样为公司提供大量信息，以保证企业获得可观的产品收益，并在该行业保持竞争力。

例如，MeadWestvaco 拥有大量的森林资源，它们为公司生产多种产品提供所需的原料——树木。管理人员需要掌握关于木材及森林准确而可靠的信息，比如，森林的现有储量如何，森林以往的生长情况如何，森林未来计划生长情况如何。MeadWestvaco 的管理人员基于这些重要问题的答案，制定未来包括树木的长期种植和采伐时间表在内的计划。

MeadWestvaco 公司拥有大量森林，它是如何获取所需要的这些资源的信息的呢？从遍布森林的抽样点收集数据是对公司拥有的树木总体进行了解的基础。为了确定抽样点，首先按照位置和树种将木材林分成三部分，利用地图和随机数，MeadWestvaco 的分析人员从每部分森林中选取1/5～1/7英亩的树木组成随机样本。MeadWestvaco 公司的林务员通过从这些抽样点收集的数据了解森林总体。

全体林务员参加数据的收集过程。他们两人一组定期收集每一抽样点中每棵树的信息。这些抽样数据被录入公司的森林永续存货（CFI）计算机系统。该系统所提供的报告汇总了大量数据的频率分布信息，对包括树木类型、现有森林储量、森林以往生产率、未来计划森林生长和储量作出统计。抽样和对抽样数据的统计汇总报告为 MeadWestvaco 有效管理森林和木材林提供了重要的保证。

本章将学习简单随机抽样和样本选择方法。另外，还要学习如何利用统计量，比如样本均值和样本比率来估计总体均值和总体比率。此外，还将介绍抽样分布的重要概念。

在第1章中，我们给出了个体、总体和样本的定义。

（1）个体是收集数据的基本单位。

（2）总体是所有感兴趣的个体的集合。

（3）样本是总体的一个子集。

我们选取样本的目的就是为了收集用于推断所需的数据，并且回答关于总体的研究问题。

下面我们从两个例子开始，这两个例子都是利用抽样来回答关于总体的一个研究问题的。

（1）得克萨斯州某个政治团体的成员正在考虑支持某候选人竞选美国参议员。政治团体领导者需要对本州登记选民中支持该名候选人的比率做一个估计。选取得克萨斯州400名登记选民组成一个样本，其中有160人对该候选人表示支持。因此，登记选民总体中支持该候选人比率的估计值为160/400＝0.4。

（2）一个轮胎制造商正在考虑生产一种新设计开发的轮胎，这种轮胎行驶的里程数超过了企业现有生产线上轮胎的里程数。为了对这种新型轮胎的平均使用寿命作出估计。制造商生产了120个这种新型轮胎组成样本用于检测，检测结果表明样本均值为36 500英里。于是，该种新型轮胎总体的平均使用寿命的估计值为36 500英里。

抽样结果提供的仅仅是相应总体特征值的估计，认识到这一点非常重要。[⊜]我们并不期望注册选民总体中恰好有40%的人支持该候选人；同样，我们也没有期望样本均值36 500英里恰好是所生产的新型轮胎总体的平均里程。原因很简单，样本只包含了总体的一部分，可以预见，会有抽样误差。利用恰当的抽样方法，抽样结果可以

⊖ 作者感谢 Edward P. Winkofsky 博士，他为“实践中的统计”提供了本案例。

⊜ 样本均值是总体均值的估计值，样本比率是总体比率的估计值。这种估计是有抽样误差的。本章给出了确定抽样误差的大小的基本原则。

给出关于总体特征的一个"好"的估计。那么，我们希望抽样结果好到什么程度呢？统计过程可以对此问题作出回答。

我们定义一些抽样的术语。从中抽取样本的总体叫作**抽样总体**（sampled population）。**抽样框**（frame）是用于抽选样本的个体清单。在第一个例子中，抽样总体是得克萨斯州所有的登记选民，抽样框是所有登记选民的清单。由于得克萨斯州所有登记选民数目是一个有限的值，因此，第一个例子是从一个有限总体进行抽样的实例。在第7.2节，我们将讨论如何从一个有限总体抽取一个简单随机样本。

在轮胎行驶里程的例子中，由于120个轮胎所组成的样本取自某一特定时点的生产过程，定义抽样总体变得尤为困难。我们可以把抽样总体看作一个由生产过程某一段特定时间段所生产的全部轮胎组成的抽象总体。从这个意义上讲，可以认为抽样总体是无限的，因此无法构建一个抽样框。在第7.2节，我们将讨论在这种情形下如何抽取一个随机样本。

本章，我们首先说明了如何采用简单随机抽样从一个有限总体中选取样本。对由连续运行过程形成的无限总体，描述如何从无限总体中选取随机样本。然后说明如何利用简单随机抽样得到的数据计算总体均值、总体标准差和总体比率的估计值。另外，我们还介绍了抽样分布的重要概念。正如将要介绍的那样，抽样分布的知识使我们可以说明样本估计值与相应总体参数的接近程度。最后一节里，我们还将讨论实际应用中除简单随机抽样之外的其他一些常用随机抽样方法。

7.1 EAI的抽样问题

EAI公司的人事部经理被分派一项任务，要求为公司2 500名管理人员制定一份简报，内容包括管理人员的平均年薪和公司中已完成公司管理培训计划的管理人员所占的比率。

2 500名管理人员构成此项研究的总体，我们可以参考公司的职员记录找到总体中每个人的年薪金额和是否完成管理培训计划。总体中2 500名管理人员的信息数据存放在名为EAI的文件内。

利用EAI的数据和第3章介绍的公式，可以计算年薪数据的总体均值和总体标准差。

$$\text{总体均值}: \mu = 51\ 800 \text{ 美元}$$
$$\text{总体标准差}: \sigma = 4\ 000 \text{ 美元}$$

数据显示，2 500名管理人员中有1 500人已经完成培训计划。

总体的数字特征称作**参数**（parameters），将总体中已完成培训计划的职员比率记做p，有$p = 1\ 500/2\ 500 = 0.6$。总体年薪均值（$\mu = 51\ 800$美元）、总体标准差（$\sigma = 4\ 000$美元）和完成培训计划的总体比率（$p = 0.6$）都是EAI管理人员总体的参数。

现在，假设我们无法从公司的数据库中获得EAI管理人员这些必要的信息。我们考虑的问题是，如果不用总体中全体2 500名管理人员，而是用一个样本的话，企业人事部经理将如何获取对这些总体参数的估计呢？假定选取30名管理人员组成一个样本，显然，这样比编写整个总体的简报要节约时间和成本。[⊖]如果人事部经理确信，30名管理人员的样本提供了关于总体中2 500名管理人员的足够信息，那么用样本比用整个总体来编写简报更好。首先，我们从考虑如何取得一个30名管理人员的样本入手，探究利用样本研究EAI问题的可能性。

7.2 抽样

本节我们描述如何选取样本。首先，我们描述从有限总体抽样，然后描述从无限总体抽样。

7.2.1 从有限总体的抽样

在从有限总体抽样时，统计学家建议采用概率抽样，因为基于概率抽样的样本可以对总体进行有效的统计推

⊖ 通常情况下，从一个样本中收集信息的成本要大大低于从总体中收集信息的成本，特别是当需要通过个人访谈来收集信息的时候。

断。若每个容量为n的样本以相同的概率被抽到，则称这种抽样方式为简单随机样本。[一]这是最简单的一种概率抽样，从容量为 N 的有限总体抽取容量为 n 的简单**随机样本**（simple radom sample）的定义如下。

简单随机样本（有限总体）

从容量为 N 的有限总体中抽取一个容量为 n 的样本，如果容量为 n 的每一个可能的样本都以相等的概率被抽出，则称该样本为简单随机样本。

自有限总体选择简单随机样本的一种抽样流程是，每次只选择一个样本点，总体中的每一个体等可能被抽到。用这种方式抽得的 n 个个体满足自有限总体的简单随机样本的定义。

为了从 EAI 管理人员组成的有限总体中抽取一个简单随机样本，我们首先为每名管理人员依次标号建立一个抽样框。比如，按照管理人员姓名在 EAI 个人信息文件中出现的顺序分别为管理人员标号 1 ~ 2 500。其次，我们查随机数表[二]（见表 7-1）。表中第一行的每个数字 6，3，2，…，都是随机数，以相同的机会发生。由于 EAI 管理人员总体的最大标号为 2 500，是四位数，我们从表中每 4 位一组选择随机数[三]。可以从表中任意一处开始，按照我们的方向机械地移动来选取随机数。我们根据表 7-1 第一行按从左到右的方向移动，得到的前 7 个四位随机数为

6 327　　1 599　　8 671　　7 445　　1 102　　1 514　　1 807

由于表中数字是随机的，因此这些四位数也是等可能的。

采用这些四位随机数，我们现在可以将总体中每一管理人员以相等的机会选入随机样本。第一个数 6 327 比 2 500大，与总体中任一管理人员的编号不符，于是舍弃；第二个数 1 599 介于 1 ~ 2 500，于是 EAI 管理人员名单中标号为 1 599 的管理人员第一个被选入随机样本。继续该过程，忽略 8 671 和 7 455，选择标号为 1 102，1 514，1 807的管理人员进入随机样本。这一过程一直继续下去，直到取得由 30 名 EAI 管理人员组成的简单随机样本。

选取简单随机样本时，在 30 名管理人员被选出之前，表中先前已经出现过的随机数可能重复出现。由于该管理人员已经被选入样本了，我们并不想将一个管理人员多次选入，所以忽略已出现过的随机数。这种选取样本的方式叫作**无放回抽样**（sampling without replacement）。如果我们选取样本时，对已经出现过的随机数仍选入样本，某些管理人员可能在样本中被两次或更多次地包括进来，则我们进行的是**放回抽样**（sampling with replacement）。抽样中，放回抽样是一种取得简单随机样本的有效途径，然而，无放回抽样更为常用。当我们提到简单随机抽样时，我们总是假定抽样是无放回的。

表 7-1　随机数表

63 271	59 986	71 744	51 102	15 141	80 714	58 683	93 108	13 554	79 945
88 547	09 896	95 436	79 115	08 303	01 041	20 030	63 754	08 459	28 364
55 957	57 243	83 865	09 911	19 761	66 535	40 102	26 646	60 147	15 702
46 276	87 453	44 790	67 122	45 573	84 358	21 625	16 999	13 385	22 782
55 363	07 449	34 835	15 290	76 616	67 191	12 777	21 861	68 689	03 263
69 393	92 785	49 902	58 447	42 048	30 378	87 618	26 933	40 640	16 281
13 186	29 431	88 190	04 588	38 733	81 290	89 541	70 290	40 113	08 243
17 726	28 652	56 836	78 351	47 327	18 518	92 222	55 201	27 340	10 493
36 520	64 465	05 550	30 157	82 242	29 520	69 753	72 602	23 756	54 935
81 628	36 100	39 254	56 835	37 636	02 421	98 063	89 641	64 953	99 337
84 649	48 968	75 215	75 498	49 539	74 240	03 4 66	49 292	36 401	45 525
63 291	11 618	12 613	75 055	43 915	26 488	41 116	64 531	56 827	30 825
70 502	53 225	03 655	05 915	37 140	57 051	48 393	91 322	25 653	06 543
06 426	24 771	59 935	49 801	11 082	66 762	94 477	02 494	88 215	27 191
20 711	55 069	29 430	70 165	45 406	78 484	31 639	52 009	18 873	96 927
41 990	70 538	77 191	25 860	55 204	73 417	83 920	69 468	74 972	38 712

[一] 关于概率抽样的其他方法将在第 7.8 节中介绍。

[二] 在本章的附录中，我们将说明如何用 Minitab，Excel 和 StatTools 从有限总体中抽取简单随机样本。

[三] 为了可读性，表中的随机数都用五位数来表示。

（续）

72 452	36 618	76 298	26 678	89 334	33 938	95 567	29 380	75 906	91 807
37 042	40 318	57 099	10 528	09 925	89 773	41 335	96 244	29 002	46 453
53 766	52 875	15 987	46 962	67 342	77 592	57 651	95 508	80 033	69 828
90 585	58 955	53 122	16 025	84 299	53 310	67 380	84 249	25 348	04 332
32 001	96 293	37 203	64 516	51 530	37 069	40 216	61 374	05 815	06 714
62 606	64 324	46 354	72 157	67 248	20 135	49 804	09 226	64 419	29 457
10 078	28 037	85 389	50 324	14 500	15 562	64 165	06 125	71 353	77 669
91 561	46 145	24 177	15 294	10 061	98 124	75 732	00 815	83 452	97 355
13 091	98 112	53 959	79 607	52 244	63 303	10 413	63 839	74 762	50 289

7.2.2　从无限总体的抽样

有时候，我们想从总体中抽取样本，但是总体容量无限大或者总体中的个体是由一个正在运行的过程产生的，从而生成的个体数量是无限的，因此无法得到总体中所有个体的清单，这是无限总体的情形。对于无限总体，由于无法构建一个包含全部个体的抽样框，因此无法抽取一个简单随机样本。对于无限总体的情形，统计学家建议抽取一个所谓的**随机样本**（random sample）。

随机样本（无限总体）

如果从一个无限总体中抽取一个容量为 n 的样本，使得下面的条件得到满足：

1. 抽取的每个个体来自同一总体；
2. 每个个体的抽取是独立的。

则称该样本是一个随机样本。

当从无限总体中抽取一个随机样本时，必须小心并仔细判断。不同情形可能需要采取不同的抽取方法。我们通过两个例子来说明条件 1“抽取的每个个体来自同一总体”和条件 2“每个个体的抽取是独立的”的含义。

在普通的质量控制的应用中，生产过程中所生产的产品数量是无限的。抽样总体是由正在运行的生产过程中生产的全部产品而不仅仅是由那些已经生产的产品组成。因为我们不可能列出生产的全部产品的清单，所以认为总体是无限的。更具体地，比如设计一条生产线用于盒装早餐麦片，早餐麦片的平均重量为每盒 24 盎司。为判断生产线是正常运行还是由于机器故障使得生产线的填充量过多或者不足，一位质量控制检验员定期从生产线上抽取 12 盒产品组成一个样本。

在这样一个生产操作中，选取一个随机样本时最关心的是条件 1“抽取的每个个体来自同一总体”是否成立。为了确保这一条件成立，必须在近似相同的时点选择产品。这样，才能避免检验员抽取的某些盒来自生产线正常运行时，而另一些产品是在生产线非正常运行，从而使得每盒的填充量过多或者不足。在诸如这样的生产过程中，设计的生产流程应确保每盒麦片的装盒是相互独立的，从而满足条件 2，即“每个个体的抽取是独立的”。在这个假定下，检验员只需关注条件“每个个体来自同一总体”是否成立即可。

从无限总体抽取随机样本的另一个例子是，考虑由到达快餐店的顾客组成的总体。假定快餐店要求一名雇员选取顾客样本，完成一个简短的调查问卷。顾客光临快餐店是一个正在进行中的过程，因此不能得到总体中所有顾客的一个名单。于是，出于应用的目的，可以将这个正在进行过程中的总体看作无限的。如果设计一种抽样方法，使得样本中的所有个体都是餐厅的顾客并且是独立选择的，那么可以得到一个随机样本。这时，雇员应该在那些进入餐馆并就餐的人中选取，以保证满足“个体来自同一总体”的条件。如果雇员选取了那些只是为了使用洗手间而进入餐馆的人组成样本，这些人并不是顾客，违背了“同一总体”的条件。因此，如果调查员是从那些来餐馆消费的人中选取样本的话，那么条件 1 可以被满足。尤为困难的是，确保独立地选取顾客。

随机样本选择方法中的第二个条件“独立选择各个个体”的用意是为了防止选择偏差。当调查员可以任意自由地选取顾客进入样本时，就可能发生选择偏倚。此时，调查员可能更愿意选择抽取某一特定年龄段的顾客进入样本，而避免从其他年龄段的顾客中抽取。再比如，五名顾客一同前来就餐，如果调查员将他们一同选入样本，也会发生

选择偏差。这样的一组顾客更易于表现出相似的特征，他们所给出的关于顾客总体的信息可能具有误导性。通过确保某一顾客的入选并不影响其他顾客的入选，则可以避免这类选择偏差。换而言之，个体（顾客）的选取是独立的。

快餐店的巨头麦当劳恰好在这种情况下实施了一次随机抽样。抽样方法以一些顾客是否持有优惠券为依据，每当一名顾客出示优惠券时，将选取下一名接受服务的顾客填写一张顾客问卷调查表。因为来到麦当劳的顾客出示优惠券是随机的，并且与其他顾客是相互独立的，因此这种抽样方式确保对顾客的选取是独立的。于是，样本满足从无限总体抽取随机样本的要求。

从无限总体抽样的情况总是与某段时间正在不断运行的过程相联系。例如，生产线上生产的零部件、实验室中反复进行的实验、银行发生的交易、技术支持中心接到的电话以及进入零售店的顾客。它们都可以看作一个从无限总体产生个体的过程。如果样本中的个体选自同一个总体，并且是独立选取的，那么样本就可以看作一个来自无限总体的随机样本。

注释和评论

1. 本节中，我们详细定义了两类样本：来自有限总体的简单随机样本和来自无限总体的随机样本。在本书余下的部分中，我们一般将这两类样本统称为随机样本或者样本。除非需要，在习题或者讨论中我们并不刻意区分样本是否是“简单”随机样本。
2. 在从有限总体的抽样调查中，统计学家使用的抽样方法是概率抽样。在概率抽样中，每个可能的样本都有一个已知的抽取概率，并且对样本中个体的抽取使用了随机方法。简单随机抽样就是这类方法当中的一种。在第7.8节，我们将介绍其他一些概率抽样方法：分层随机抽样、整群抽样和系统抽样。在简单随机抽样中，“简单”这个词是为了保证这是一种概率抽样，每个容量为 n 的样本都以相同的概率被选取。
3. 自一个容量为 N 的有限总体，选取容量为 n 的简单随机样本，有

$$\frac{N!}{n!(N-n)!}$$

种不同的选法。公式中，$N!$ 和 $n!$ 是第4章中所讲过的阶乘运算。在EAI问题中，$N=2\,500$，$n=30$，从上式可知，由30名EAI管理人员组成的不同的简单随机样本的数目大约为 2.75×10^{69} 个。

练 习

方法

2. 假定有限总体中有350个个体，用下面五位随机数的后3位（比如：601，022，448，…），确定被选入简单随机样本的前4个个体。

98 601　73 022　83 448　34 229　27 553　84 147
93 289　14 209

应用

4. 纽约证券交易所（NYSE）2006年3月6日10支最活跃的股票如下（《华尔街日报》，2006年3月7日）。

美国电话电报	美国辉瑞
美国 Lucent 公司	美国德州仪器
美国北电	美国通用电气
美国 Qwest	美国 iShrMSJpn
美国南方贝尔	美国 LSI Logic

交易当局决定从这些股票中选取3只组成的一个样本，对交易业务进行调查。

a. 从表7-1第6列的第一个随机数开始按列向下，选择3只股票构成一个简单随机样本提供给交易当局。

b. 根据本节注释3中的信息，从10只股票组成的清单中能选出多少个不同的样本容量为3的简单随机样本？

6. 由美国人口统计局出版的《全美县和市的数据》一书中列示了全美3 139个县（市）的信息。假定一项国内研究中，要从中随机选取30个县（市）收集数据。利用表7-1最后一列的四位随机数做为选择样本时头5个县（市）的编号。忽略头一位数，从四位随机数9 945，8 364，5 702等开始。

8. 下表是道琼斯工业平均指数中的30种成分股（《巴伦周刊》，2012年6月30日）。假定你想要从中选取6家公司组成一个样本，对管理实践进行深入研究。利用表7-1第9列中每行的前两位数字，从中选取6家公司组成一个简单随机样本。

1. 3M公司	16. 家得宝
2. 美国电话电报公司	17. IBM
3. 美国铝业	18. 英特尔
4. 美国运通公司	19. 强生制药有限公司
5. 美国银行	20. 卡夫食品公司
6. 波音公司	21. 麦当劳
7. 卡特彼勒公司	22. 默克制药
8. 雪佛龙公司	23. 微软
9. 思科公司	24. 摩根大通银行
10. 可口可乐公司	25. 辉瑞制药
11. 迪士尼	26. 宝洁公司
12. 杜邦公司	27. Travelers 保险公司
13. 埃克森美孚公司	28. 联合科技
14. 通用电气	29. 威瑞森公司
15. 惠普	30. 沃尔玛

10. 说明下列情形下，哪些是自有限总体的抽样，哪些是自无限总体的抽样。当抽样总体是有限的时候，说明你是如何建立抽样框的。
   a. 抽取纽约州注册驾驶员的一个样本。
   b. 抽取 Breakfast Choice 公司生产的盒装麦片的一个样本。
   c. 抽取某工作日通过金门大桥的汽车的一个样本。
   d. 抽取印第安纳大学选修统计课的学生的一个样本。
   e. 抽取某邮购业务公司处理的订单的一个样本。

7.3 点估计

我们已经描述了如何选取一个简单随机样本，现在我们回到 EAI 问题。假定已选取了一个由 30 名管理人员组成的简单随机样本，他们相应的年薪及参加管理培训计划的数据如表 7-2 所示。符号 x_1，x_2，…代表样本中第一名管理人员的年薪、第二名管理人员的年薪……在管理培训计划这一栏，已参加过管理培训计划的人员用“是”表示。

为了估计总体参数，计算相应的样本特征——**样本统计量**（sample statistic）。例如，为了估计 EAI 管理人员年薪的总体均值 μ 和总体标准差 σ，我们用表 7-2 的数据计算相应的样本统计量：样本均值和样本标准差 s。根据第 3 章中样本均值和样本标准差的公式，得样本均值为

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1\,554\,420}{30} = 51\,814(\text{美元})$$

样本标准差为

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{325\,009\,260}{29}} = 3\,348(\text{美元})$$

表 7-2 30 名 EAI 管理人员组成的简单随机样本中年薪和培训计划情况的数据

年薪（美元）	是否参加管理培训计划	年薪（美元）	是否参加管理培训计划	年薪（美元）	是否参加管理培训计划
x_1 = 49 094. 30	是	x_{11} = 45 922. 60	是	x_{21} = 45 120. 90	是
x_2 = 53 263. 90	是	x_{12} = 57 268. 40	否	x_{22} = 51 753. 00	是
x_3 = 49 643. 50	是	x_{13} = 55 688. 80	是	x_{23} = 54 391. 80	否
x_4 = 49 894. 90	是	x_{14} = 51 564. 70	否	x_{24} = 50 164. 20	否
x_5 = 47 621. 60	否	x_{15} = 56 188. 20	否	x_{25} = 52 973. 60	否
x_6 = 55 924. 00	是	x_{16} = 51 766. 00	是	x_{26} = 50 241. 30	否
x_7 = 49 092. 30	是	x_{17} = 52 541. 30	否	x_{27} = 52 793. 90	否
x_8 = 51 404. 40	是	x_{18} = 44 980. 00	是	x_{28} = 50 979. 40	是
x_9 = 50 957. 70	是	x_{19} = 51 932. 60	是	x_{29} = 55 860. 90	是
x_{10} = 55 109. 70	是	x_{20} = 52 973. 00	是	x_{30} = 57 309. 10	否

为了估计总体中完成管理培训计划的管理人员所占比率 p，我们使用与之对应的样本比率 $\bar{p}$。令 x 表示样本中完成管理培训计划的管理人员的人数。表 7-2 中数据显示，$x = 19$。样本容量 $n = 30$，从而，样本比率为

$$\bar{p} = \frac{x}{n} = \frac{19}{30} = 0.63$$

通过上述计算，我们完成了称为**点估计**的统计过程。我们称样本均值 $\bar{x}$ 为总体均值 μ 的**点估计量**（point esti-

mator），称 s 为总体标准差 σ 的点估计量，称样本比率 $\bar{p}$ 为总体比率 p 的点估计量。$\bar{x}$、s 和 $\bar{p}$ 的数值称为**点估计值**（point estimate）。于是，对于由表 7-2 所列出的 30 名 EAI 管理人员组成的简单随机样本，51 814 美元是 μ 的点估计值，3 348 美元是 σ 的点估计值，0.63 是 p 的点估计值。表 7-3 总结了这些样本结果，并且将点估计值与总体参数的真值做了对比。

表 7-3 30 名管理人员组成的简单随机样本中点估计值的小结

总体参数	参数值	点估计量	点估计值
μ——年薪的总体均值	51 800 美元	$\bar{x}$——年薪的样本均值	51 814 美元
σ——年薪的总体标准差	4 000 美元	s——年薪的样本标准差	3 348 美元
p——已完成管理培训计划的总体比率	0.60	$\bar{p}$——已完成管理培训计划的样本比率	0.63

从表 7-3 可见，点估计值与总体参数的真值在某种程度上是有差异的。这个差异是可以预期的，因为在进行点估计时用的是来自总体的样本而不是对整个总体的普查。在下一章中，我们将说明如何构造区间估计以便提供关于点估计值与总体参数差异大小的信息。

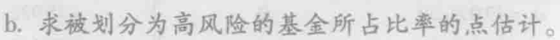
7.3.1 应用中的建议

本书的其余部分大都是关于统计推断的内容，点估计是统计推断的一种形式。我们使用样本统计量对总体参数进行推断。当根据样本对总体进行推断时，抽样总体和目标总体之间具有密切的对应是非常重要的。**目标总体**（target population）是指我们想要推断的总体，抽样总体是指实际抽取样本的总体。本节我们描述了从 EAI 管理者总体中抽取简单随机样本的过程，并且对这一总体的特征进行点估计，因此抽样总体与目标总体是同一个，这正是我们所期望的。但是，在其他情形下，并不总是容易得到具有密切对应的抽样总体和目标总体。

比如在游乐园的例子中，游乐园选取游客组成一个样本，了解诸如游客的年龄和游园时间这样一些特征。假定所有样本中的个体都选自同一天，而这天仅对某个企业的员工开放。那么，抽样总体应该由该企业的雇员及其家庭成员构成。如果我们想要推断的目标总体是某年夏季时公园的游客，那么我们面对的抽样总体与目标总体存在显著差异。在这种情形下，我们对所做的点估计的有效性产生质疑。公园管理者必须明确取自某一天的样本是否可以看作目标总体的一个代表。

总之，当利用样本去推断总体时，我们应该确保所设计的研究中抽样总体与目标总体是高度一致的。良好的判断是合理应用统计方法的基础。

练习



方法

12. 就某一问题对由 150 人组成的样本进行调查，结果为：75 人表示同意，55 人表示反对，20 人表示弃权。
    a. 总体中表示同意的人的比率的点估计值是多少？
    b. 总体中表示反对的人的比率的点估计值是多少？

应用

14. 《商业周刊》公布了 283 支证券投资基金的信息（《商业周刊》，2004 年 1 月 26 日）。这些基金中的 40 支组成一个样本，存放在数据集 MutualFund 中。利用该数据集回答下列问题。
    a. 求《商业周刊》证券投资基金中收费基金所占比率的点估计。
    b. 求被划分为高风险的基金所占比率的点估计。
    c. 求低于平均风险率的基金所占比率的点估计。
16. 从美国 50 岁及以上的成人中选取 426 人组成一个样本，询问他们"2012 年总统选举中，你最关心的问题是什么？"（《美国退休人员协会公告》，2012 年 3 月）。
    a. 研究中的抽样总体是什么？
    b. 受访者中有 350 人认为社会医疗保险是最为重要的。试估计美国 50 岁及以上的成人中认为社会医疗保险最为重要的人所占的比率。
    c. 74% 的受访者认为教育是最为重要的。试估计认为教育最为重要的受访者有多少人？
    d. 受访者中有 354 人认为增加就业是最为重要的。

试估计美国50岁及以上的成人中认为增加就业最为重要的人所占的比率。

e. 在（b）和（d）的推断中，目标总体是什么？与（a）中的抽样总体相同吗？假定你后来获悉，样本局限在美国退休人员协会成员中。你仍然认为（b）和（d）中的推断有效吗？为什么？

7.4 抽样分布简介

在上一节，我们说样本均值 $\bar{x}$ 是总体均值 μ 的点估计，样本比率 $\bar{p}$ 是总体比率 p 的点估计。表7-2中的30名EAI管理人员组成一个简单随机样本，μ的点估计值 $\bar{x}=51\,814$ 美元，p 的点估计值 $\bar{p}=0.63$。假定我们选取另一个由30名EAI管理人员组成的简单随机样本，得到如下点估计值：

$$\text{样本均值}:\bar{x}=52\,670\text{ 美元}$$

$$\text{样本比率}:\bar{p}=0.70$$

注意，得到了不同的 $\bar{x}$ 和 $\bar{p}$ 的值。的确，不能预期由30名EAI管理人员组成的第二个简单随机样本与第一个简单随机样本所得到的点估计值相同。

现在，假定将选取30名管理人员组成一个简单随机样本的过程一而再再而三地进行下去，每次都计算 $\bar{x}$ 和 $\bar{p}$ 的值。表7-4是500个这样的简单随机样本所得到的部分结果。表7-5给出的是500个 $\bar{x}$ 值的频数及频率分布。图7-1是 $\bar{x}$ 的相对频率直方图。

表7-4　由30名EAI管理人员组成的500个简单随机样本的 $\bar{x}$ 和 $\bar{p}$

样本编号	样本均值（$\bar{x}$）	比率（$\bar{p}$）	样本编号	样本均值（$\bar{x}$）	比率（$\bar{p}$）
1	51 814	0.63	4	51 588	0.53
2	52 670	0.70	⋮	⋮	⋮
3	51 780	0.67	500	51 752	0.50

表7-5　由30名EAI管理人员组成的500个简单随机样本的 $\bar{x}$ 的频数和相对频率

年薪均值（美元）	频数	相对频率	年薪均值（美元）	频数	相对频率
49 500.00～49 999.99	2	0.004	52 000.00～52 499.99	110	0.220
50 000.00～50 499.99	16	0.032	52 500.00～52 999.99	54	0.108
50 500.00～50 999.99	52	0.104	53 000.00～53 499.99	26	0.052
51 000.00～51 499.99	101	0.202	53 500.00～53 999.99	6	0.012
51 500.00～51 999.99	133	0.266	合计	500	1.000

在第5章中，我们将随机变量定义为对试验结果的数值描述。如果我们将抽取一个简单随机样本的过程看作一个试验，则样本均值 $\bar{x}$ 就是对试验结果的一个数值描述。从而，样本均值 $\bar{x}$ 是一个随机变量。因此，就像其他随机变量一样，$\bar{x}$ 有均值或数学期望、标准差和概率分布。在不同的简单随机样本中，$\bar{x}$ 的取值也有各种可能的结果，我们称 $\bar{x}$ 的概率分布为 $\bar{x}$ 的**抽样分布**（sample distribution）。抽样分布的知识和性质使我们能够对样本均值 $\bar{x}$ 与总体均值 μ 的接近程度做一个概率度量。[⊖]

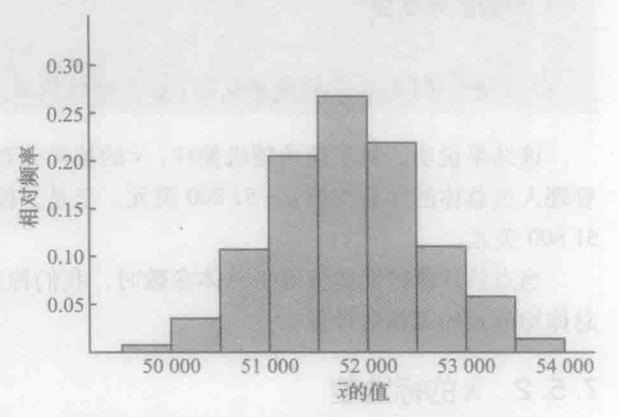


图7-1　500个样本容量为30的简单随机样本的 $\bar{x}$ 值的相对频率直方图

让我们回到图7-1，我们需要列举出所有可能的由30名管理人员组成的样本，并计算每个样本均值从而确定 $\bar{x}$ 的抽样分布。500个 $\bar{x}$ 值的直方图是对抽样分布的一个近似。从这个近似可见，分布形状是钟形的。我们还注意到，绝大部分 $\bar{x}$ 的取值聚集在一起，500个 $\bar{x}$ 值的均值在总体均值 $\mu=51\,800$

⊖ 在以下各章资料里，对内容的理解能力，很大程度上依赖于对本章中所介绍的抽样分布的理解和使用能力。

美元附近。在下一节，我们将更全面地描述 $\bar{x}$ 的抽样分布的性质。

样本比率 $\bar{p}$ 的500个值的相对频率直方图汇总在图7-2中。与 $\bar{x}$ 的情形一样，$\bar{p}$ 是一个随机变量。如果从总体中将容量为30的所有可能的样本都选出，并且计算每个样本的 $\bar{p}$ 值，所得到的概率分布叫 $\bar{p}$ 的抽样分布。图7-2是500个样本的样本比率的相对频率直方图，它给出了 $\bar{p}$ 的抽样分布的一般外形特点。

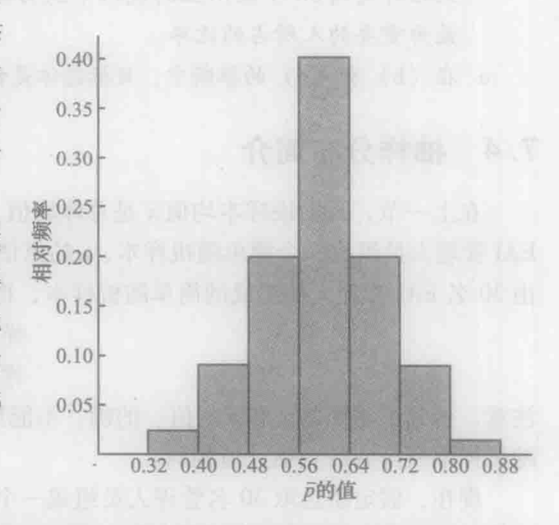


图7-2 500个样本容量为30的简单随机样本的 $\bar{p}$ 值的相对频率直方图

在实践中，我们只从总体中抽取一个简单随机样本。在本节我们将抽样过程简单地重复进行了500次，仅仅是为了说明可能取得多种不同的样本，而且不同的样本得到的样本统计量 $\bar{x}$ 和 $\bar{p}$ 的值也是不尽相同的。任何特定的样本统计量的概率分布称为该统计量的抽样分布。我们将在第7.5节说明 $\bar{x}$ 的抽样分布的特征，在第7.6节说明 $\bar{p}$ 的抽样分布的特征。

7.5 $\bar{x}$ 的抽样分布

在前一节中，我们说样本均值 $\bar{x}$ 是一个随机变量，称它的概率分布为 $\bar{x}$ 的抽样分布。

$\bar{x}$ 的抽样分布

$\bar{x}$ 的抽样分布是样本均值 $\bar{x}$ 的所有可能值的概率分布。

本节描述 $\bar{x}$ 的抽样分布的性质。与我们研究的其他概率分布一样，$\bar{x}$ 的抽样分布有均值或数学期望、标准差以及形状或形态特征。我们首先考虑 $\bar{x}$ 所有可能值的均值，即 $\bar{x}$ 的数学期望。

7.5.1 $\bar{x}$ 的数学期望

在 *EAI* 抽样问题中，我们看到不同的简单随机样本得出的样本均值 $\bar{x}$ 是不同的。因为随机变量 $\bar{x}$ 可能有许多不同的值，所以我们关心的是由大量简单随机样本产生的 $\bar{x}$ 的所有可能值的均值。随机变量 $\bar{x}$ 的均值是 $\bar{x}$ 的数学期望。令 $E(\bar{x})$ 表示 $\bar{x}$ 的数学期望，μ 表示我们抽取简单随机样本时总体的均值。对简单随机抽样，可以证明 $E(\bar{x})$ 与 μ 相等。

$\bar{x}$ 的数学期望[⊖]

$$E(\bar{x}) = \mu \tag{7-1}$$

式中，$E(\bar{x})$ 为 $\bar{x}$ 的数学期望；μ 为总体均值。

该结果说明，对于简单随机抽样，$\bar{x}$ 的抽样分布的均值或数学期望等于总体均值。在第7.1节，我们看到EAI管理人员总体的年薪均值 μ = 51 800美元。于是，根据式（7-1），EAI研究中样本均值所有可能值的均值也等于51 800美元。

当点估计量的期望值等于总体参数时，我们称这个点估计量是**无偏的**（unbiased）。由式（7-1）可见，$\bar{x}$ 是总体均值 μ 的无偏估计量。

7.5.2 $\bar{x}$ 的标准差

采用以下记号，我们给出 $\bar{x}$ 的抽样分布的标准差的定义。

⊖ $\bar{x}$ 的数学期望等于抽样总体的均值。

$\sigma_{\bar{x}}$——$\bar{x}$ 的标准差

σ——总体标准差

n——样本容量

N——总体容量

可以证明，$\bar{x}$ 的标准差公式与总体是否有限有关。下面给出 $\bar{x}$ 的标准差的两个公式。

$\bar{x}$ 的标准差

有限总体　　　　无限总体

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{7-2}$$

比较式（7-2）中的两个公式，我们看到对有限总体需要系数 $\sqrt{(N-n)/(N-1)}$，而无限总体则不需要该系数。通常称该系数为**有限总体修正系数**（finite population correction factor）。在许多实际抽样中，我们发现虽然总体是有限的，但容量很"大"，相对而言样本容量很"小"，这时有限总体修正系数 $\sqrt{(N-n)/(N-1)}$ 趋近于1，可以忽略有限总体 $\bar{x}$ 的标准差与无限总体 $\bar{x}$ 的标准差之间的差别。于是，虽然总体有限，但可以用 $\sigma_{\bar{x}}=\sigma/\sqrt{n}$ 作为 $\bar{x}$ 的标准差的一个很好的近似。由此观察可以得出下面计算 $\bar{x}$ 的标准差的一般指导方针或经验法则。

计算 $\bar{x}$ 的标准差的公式

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{7-3}$$

当（1）总体是无限的；或（2）总体是有限的，但样本容量不大于总体容量的5%，即 $n/N \leqslant 0.05$。

当 $n/N>0.05$ 时，使用式（7-2）中有限总体的公式计算 $\sigma_{\bar{x}}$。除非特别说明，本书中我们假定总体容量足够"大"，使得 $n/N\leqslant 0.05$，可以用式（7-3）计算 $\sigma_{\bar{x}}$。[⊖]

为了计算 $\sigma_{\bar{x}}$，必须已知总体标准差 σ。为了更加强调 $\sigma_{\bar{x}}$ 与 σ 的不同，我们称 $\bar{x}$ 的标准差 $\sigma_{\bar{x}}$ 为均值的**标准误差**（standard error）[⊜]。一般地，标准误差指的是点估计量的标准差。随后我们将看到，均值的标准误差有助于确定样本均值与总体均值的偏离程度。回到例子 EAI 的问题中，对于 30 名 EAI 管理人员组成的简单随机样本，计算均值的标准误差。

在第 7.1 节，我们看到总体由 2 500 名 EAI 管理人员组成，年薪的标准差 $\sigma=4\,000$ 美元。这时，总体是有限的，$N=2\,500$。当样本容量为30 时，有 $n/N=30/2\,500=0.12$。因为样本容量小于总体容量的5%，所以可以忽略有限总体修正系数，采用式（7-3）计算标准误差。

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{400}{\sqrt{30}} = 730.3$$

7.5.3 $\bar{x}$ 的抽样分布的形式

前面关于 $\bar{x}$ 的抽样分布的数学期望和标准差的结论，适用于所有的总体。$\bar{x}$ 的概率分布的形式或者形态是确定 $\bar{x}$ 抽样分布特征的最后一步。考虑以下两种情形：（1）总体服从正态分布；（2）总体不服从正态分布。

总体服从正态分布　在许多情况下，有理由假设我们抽取随机样本的总体服从正态分布或近似服从正态分布。当总体服从正态分布时，在任何样本容量下 $\bar{x}$ 的抽样分布都是正态分布。

总体不服从正态分布　当我们抽取随机样本的总体不服从正态分布时，**中心极限定理**（central limit theorem）

⊖ 问题21 中表明，当 $n/N\leqslant 0.05$ 时，有限总体修正系数对 $\sigma_{\bar{x}}$ 的值几乎不产生影响。

⊜ 在统计推断中，标准误差指的是点估计量的标准差。

能帮助我们确定 $\bar{x}$ 的抽样分布的形状。中心极限定理在 $\bar{x}$ 抽样分布中的应用如下所述

中心极限定理

从总体中抽取容量为 n 的简单随机样本，当样本容量很大时，样本均值 $\bar{x}$ 的抽样分布近似服从正态概率分布。

图 7-3 说明中心极限定理对于三个不同总体的作用，每列对应于一个总体。最上面的图显示总体都是非正态的。总体Ⅰ服从均匀分布。总体Ⅱ经常被称为兔耳形分布，它是对称的，但是绝大部分值落在分布的尾端。总体Ⅲ的形状与指数分布相似，是右偏的。

图 7-3 最后三行中的图给出了当样本容量分别为 $n=2$，$n=5$ 和 $n=30$ 时抽样分布的形状。当样本容量为 2 时，我们看到每个抽样分布的外形开始呈现出与总体分布不同的外形。当样本容量为 5 时，我们看到与总体Ⅰ和总体Ⅱ所对应的抽样分布的形状都开始看上去与正态分布的形状类似。与总体Ⅲ所对应的抽样分布的形态虽然开始看上去与正态分布的形状类似，但仍呈现右偏。最后，当样本容量为 30 时，我们看到三个抽样分布的形态都近似于正态分布。

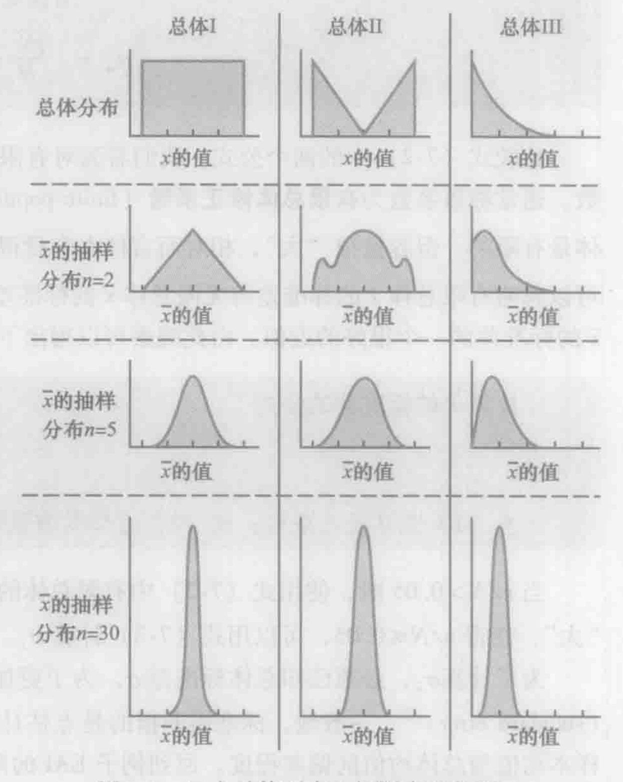


图 7-3 三个总体下中心极限定理的图示

从应用者的角度看，样本容量应该达到多大时，我们才可以应用中心极限定理？样本容量应该达到多大时，才能够假定抽样分布的形态是近似正态的呢？统计研究人员通过分析各种总体不同样本容量下 $\bar{x}$ 的抽样分布，对该问题进行了研究。在一般的统计实践中，对于大多数应用，假定当样本容量大于或等于 30 时，$\bar{x}$ 的抽样分布可用正态分布近似。当总体是严重偏态或者出现异常点时，可能需要样本容量达到 50。最后，当总体为离散型时，正态近似中所需样本容量一般依赖于总体的比例。在第 7.6 节中研究 $\bar{p}$ 的抽样分布时，我们会对该问题进行更深入的讨论。

7.5.4 EAI 问题中 $\bar{x}$ 的抽样分布

让我们回到 EAI 问题中，先前给出了 $E(\bar{x})=51\,800$，$\sigma_{\bar{x}}=730.3$。从这个角度上讲，我们没有关于总体分布的任何信息，总体可能是也可能不是正态分布。如果总体分布是正态分布，则 $\bar{x}$ 的抽样分布是正态分布；如果总体分布不是正态分布，由于我们使用的是 30 名管理人员组成的一个简单随机样本，根据中心极限定理，我们得到结论：$\bar{x}$ 的抽样分布可以用正态分布近似。在任何一种情况下，我们都很容易地得到结论：$\bar{x}$ 的抽样分布可以用图 7-4 中的正态分布描述。

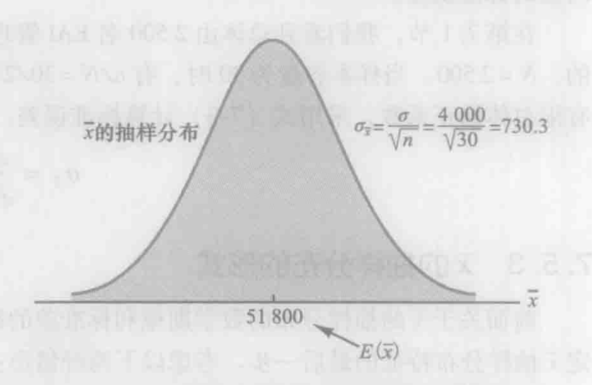


图 7-4 30 名管理人员的简单随机样本年薪均值 $\bar{x}$ 的抽样分布

7.5.5 $\bar{x}$ 的抽样分布的实际值

当抽取一个简单随机样本，用样本均值 $\bar{x}$ 的值估计总体均值 μ 时，我们不能希望样本均值恰好与总体均值相等。我们对 $\bar{x}$ 的抽样分布感兴趣的实际原因是，它可以用来提供样本均值 $\bar{x}$ 的值和总体均值 μ 的值之间差异的概率信息。为了证明这一用途，我们仍回到 EAI 问题中去。

假定人事部经理认为当样本均值在总体均值附近 ±500 美元以内时，样本均值是总体均值的一个可接受的估计值。当然，并不能确保样本均值在总体均值附近 ±500 美元以内。事实上，表 7-5 和图 7-1 显示 500 个样本均值

中有一部分与总体均值的差异是在2 000美元以上的。以概率的术语来讲，人事部经理关心的是如下问题：根据30名EAI管理人员组成的简单随机样本，得到的样本均值在总体均值附近±500美元以内的概率有多大?

我们已经确定了$\bar{x}$的抽样分布的性质（见图7-4)，可以用该分布回答概率问题。$\bar{x}$的抽样分布如图7-5所示。当总体均值为51 800美元，人事部经理关心的是样本均值介于51 300~52 300的概率。图7-5所示的抽样分布中阴影部分的面积恰好给出了这个概率值。由于抽样分布是正态分布，均值为51 800美元，标准差为730.3美元，我们用标准正态概率表来查找此概率或面积。

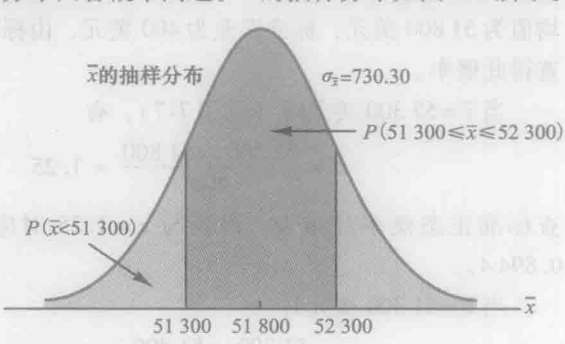


图7-5 样本均值在总体均值附近±500美元以内的概率

我们首先计算区间右端点（52 300）对应的z值，并通过查表得到曲线下方该点左侧区域（左尾部）的面积。然后，我们再计算区间左端点（51 300）对应的z值，并通过查表得到曲线下方该点左侧区域（左尾部）的面积。从后者中减去前者，即得到所求概率。

当$\bar{x}=52\ 300$美元时，有

$$z=\frac{52\ 300-51\ 800}{730.3}=0.68$$

查标准正态概率分布表，得到累计概率（$z=0.68$左侧的面积）等于0.751 7。

当$\bar{x}=51\ 300$美元时，有

$$z=\frac{51\ 300-51\ 800}{730.3}=-0.68$$

曲线下$z=-0.68$左侧的面积为0.248 3。于是，$P\{51\ 300\leqslant\bar{x}\leqslant52\ 300\}=P\{z\leqslant0.68\}-P\{z\leqslant-0.68\}=0.751\ 7-0.248\ 3=0.503\ 4$。

上述计算说明，由30名EAI管理人员组成的一个简单随机样本中，以0.503 4的可靠性保证样本均值$\bar{x}$在总体均值附近±500美元以内，即样本均值$\bar{x}$与总体均值$\mu=51\ 800$美差异大于500美元的概率为$1-0.503\ 4=0.496\ 6$。换而言之，由30名EAI管理人员组成的简单随机样本中，大约有50:50的机会使得样本均值在所允许的500美元范围以内，也许应该考虑增加样本容量。我们通过考虑样本容量与抽样分布的关系来研究这一可能性[⊖]。

7.5.6 样本容量与$\bar{x}$的抽样分布的关系

假定在EAI抽样问题中，我们最初选取的不是30名EAI管理人员而是100名EAI管理人员组成的一个简单随机样本。直觉上，似乎样本容量越大提供的数据越多，用$n=100$时的样本均值估计总体均值应该比用$n=30$的样本均值估计总体均值更好。为了说明更好的程度，我们考虑样本容量与$\bar{x}$的抽样分布的关系。

首先注意到$E(\bar{x})=\mu$与样本容量无关，$\bar{x}$所有可能值的均值等于总体均值，与样本容量n无关。然而，均值的标准误差$\sigma_{\bar{x}}=\sigma/\sqrt{n}$与样本容量的平方根有关。当样本容量增加时，均值的标准误差$\sigma_{\bar{x}}$减小。当$n=30$时，EAI问题中均值的标准误差为730.3，然而当样本容量增加到$n=100$时，均值的标准差减少到

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}=\frac{4\ 000}{\sqrt{100}}=400$$

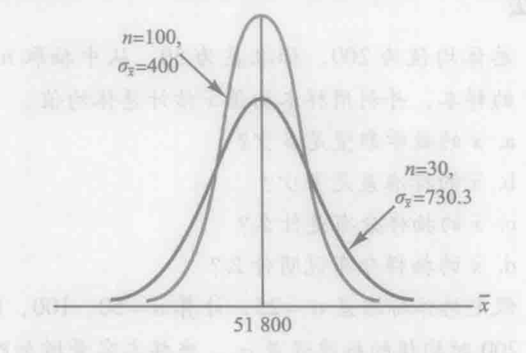


图7-6 $n=30$和$n=100$时EAI管理人员组成的简单随机样本中$\bar{x}$抽样分布的对比

$n=30$和$n=100$时$\bar{x}$的抽样分布如图7-6所示。由于$n=100$时

⊖ $\bar{x}$的抽样分布可用于提供样本均值$\bar{x}$与总体均值μ的接近程度的概率信息。

的抽样分布有更小的标准误差，$\bar{x}$的值具有更小的变异性，比$n=30$时$\bar{x}$的值更接近总体均值。

对于由100名EAI管理人员组成的一个简单随机样本，此时$n=100$，利用$\bar{x}$的抽样分布可以计算样本均值在总体均值附近±500美元以内的概率。由于抽样分布是正态的，均值为51 800美元，标准误差为400美元，由标准正态概率表可查得此概率。

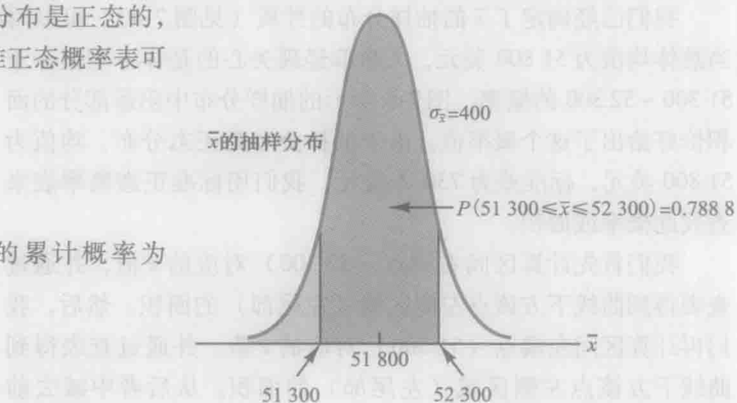


图7-7 样本容量$n=100$时EAI管理人员组成的简单随机样本的样本均值落在总体均值附近±500美元能内的概率

当$\bar{x}=52\,300$美元时（见图7-7），有

$$z=\frac{52\,300-51\,800}{400}=1.25$$

查标准正态概率分布表，得到与$z=1.25$对应的累计概率为0.894 4。

当$\bar{x}=51\,300$美元时

$$z=\frac{51\,300-51\,800}{400}=-1.25$$

与$z=-1.25$相对应的累计概率为0.105 6。于是，$P\{51\,300\leqslant\bar{x}\leqslant52\,300\}=P\{z\leqslant1.25\}-P\{z\leqslant-1.25\}=0.894\,4-0.105\,6=0.788\,8$。因而，随着样本容量从30名EAI管理人员增加到100名，样本均值在总体均值附近±500美元以内的概率从0.503 4增加到0.788 8。

该讨论的重点是随着样本容量的增加，均值的标准误差在减少。结果，样本容量越大，样本均值落在总体均值附近某一特定范围内的概率也越大。

注释和评论

1. 我们基于总体均值$\mu=51\,800$美元和总体标准差$\sigma=4\,000$美元是已知的事实，在EAI问题中得到$\bar{x}$的抽样分布时。然而，用于确定$\bar{x}$的抽样分布所需要的总体均值μ和总体标准差σ的值通常是未知的。在第8章，我们将给出当μ和σ未知时如何利用样本均值$\bar{x}$和样本标准差s。
2. 中心极限定理的理论证明中，要求样本中的观察值是独立的。自无限总体和有限总体的有放回抽样所得到的样本，都是满足该条件的。虽然中心极限定理没有直接要求自有限总体的抽样是有放回的，但是一般统计应用中只有当总体容量较大的情形下才使用中心极限定理的结果。

练 习

方法

18. 总体均值为200、标准差为50。从中抽取$n=100$的样本，并利用样本均值$\bar{x}$估计总体均值。
    a. $\bar{x}$的数学期望是多少？
    b. $\bar{x}$的标准差是多少？
    c. $\bar{x}$的抽样分布是什么？
    d. $\bar{x}$的抽样分布说明什么？

20. 假定总体标准差$\sigma=25$，计算$n=50$，100，150和200时均值的标准误差$\sigma_{\bar{x}}$。当样本容量增加时，均值的标准误差如何变化？

应用

22. EAI抽样问题中，假定由60名管理人员组成一个简单随机样本。
    a. 当简单随机样本的样本容量为60时，简述$\bar{x}$的抽样分布。
    b. 当简单随机样本的样本容量为120时，$\bar{x}$的抽样分布如何变化？
    c. 随着样本容量的增加，你认为$\bar{x}$的抽样分布通常发生怎样的变化？这看上去合乎逻辑吗？为什么？

24. 据《巴伦周刊》报道，个人失业的平均时间为17.5周（《巴伦周刊》，2008年2月18日）。假定由全体失业者组成的总体中，失业时间的总体均值是17.5周，总体标准差为4周。假定你想要选取50名失业人员组成一个随机样本进行进一步的研究。

a. 求 $\bar{x}$ 的抽样分布，其中 $\bar{x}$ 是50名失业人员的样本均值。

b. 50名失业人员组成的简单随机样本中，样本均值落在总体均值附近 ±1 周以内的概率是多少？

c. 50名失业人员组成的简单随机样本中，样本均值落在总体均值附近 ±1/2 周以内的概率是多少？

26. 汽车保险费的年均值为939美元（*CNBC*，2006年2月23日）。假定标准差 $\sigma=245$ 美元。

a. 由汽车保险政策组成一个简单随机样本，对以下样本容量，分别计算样本均值在总体均值附近 ±25 美元以内的概率是多少？$n=30$，50，100和400。

b. 当试图估计总体均值时，大样本的好处是什么？

28. 男性高尔夫球员的平均得分为95，女性高尔夫球员的平均得分为106（《高尔夫文摘》，2006年春）。以这些值作为男性球员和女性球员的总体均值，并假定二者的总体标准差均为 $\sigma=14$。选取30名男球员组成一个样本，选取45名女球员组成另一个样本。

a. 求男性球员的 $\bar{x}$ 的抽样分布。

b. 对于男性球员样本，样本均值在总体均值附近 ±3 以内的概率是多少？

c. 对于女性球员样本，样本均值在总体均值 ±3 以内的概率是多少？

d. 在b和c中，哪种情形下样本均值在总体均值附近 ±3 以内的概率更大？为什么？

30. 4 000名雇员组成一个总体，从中选取40名雇员组成一个简单随机样本，用于估计平均年龄。

a. 在计算均值的标准误差时，你是否要用有限总体修正系数？为什么？

b. 若总体标准差 $\sigma=8.2$ 年，计算均值的标准误差。计算中分别使用有限总体修正系数和不用有限总体修正系数这两种方法。当 $n/N\leq 0.05$ 时，忽略有限总体修正系数的理由是什么？

c. 雇员年龄的样本均值落在总体均值附近 ±2 年以内的概率为多少？

7.6　$\bar{p}$ 的抽样分布

样本比率 $\bar{p}$ 是总体比率 p 的点估计。样本比率的计算公式为

$$\bar{p}=\frac{x}{n}$$

式中，x 代表样本中具有感兴趣特征的个体的个数；n 代表样本容量。

正如第7.4节中所述，样本比率 $\bar{p}$ 是一个随机变量，称它的概率分布为样本比率 $\bar{p}$ 的抽样分布。

> **$\bar{p}$ 的抽样分布**
>
> $\bar{p}$ 的抽样分布是样本比率 $\bar{p}$ 的所有可能值的概率分布。

为了确定样本比率 $\bar{p}$ 与总体比率 p 的接近程度，我们需要了解 $\bar{p}$ 的抽样分布的性质：$\bar{p}$ 的数学期望、$\bar{p}$ 的标准差以及 $\bar{p}$ 的抽样分布的形状或形态。

7.6.1　$\bar{p}$ 的数学期望

$\bar{p}$ 的数学期望是 $\bar{p}$ 的所有可能值的均值，它与总体比率 p 相等。

> **$\bar{p}$ 的数学期望**
>
> $$E(\bar{p})=p \tag{7-4}$$
>
> 式中，$E(\bar{p})$ 为 $\bar{p}$ 的数学期望；p 为总体比率。

由于 $E(\bar{p})=p$，所以 $\bar{p}$ 是 p 的无偏估计量。在第7.1节中，我们已经注意到，EAI的总体比率 $p=0.6$，其中 p 表示管理人员中参加公司管理培训计划的总体比率。从而，在EAI抽样问题中，$\bar{p}$ 的数学期望为0.6。

7.6.2　$\bar{p}$ 的标准差

与样本均值 $\bar{x}$ 的标准差一样，我们发现 $\bar{p}$ 的标准差与总体是有限还是无限有关。下面给出计算 $\bar{p}$ 的标准差的

两个公式。

$\bar{p}$ 的标准差

有限总体 无限总体

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{7-5}$$

比较式（7-5）中的公式，可见不同之处仅仅在于是否使用有限总体修正系数 $\sqrt{(N-n)/(N-1)}$。

与样本均值 $\bar{x}$ 的情形一样，若有限总体的总体容量相对于样本容量足够大，有限总体与无限总体在表达式上的不同可以忽略不计。我们遵循与样本均值 $\bar{x}$ 的标准差相同的经验法则，即：如果总体是有限的并且 $n/N \leqslant 0.05$ 时，我们采用公式 $\sigma_{\bar{p}} = \sqrt{p(1-p)/n}$；然而，如果总体是有限的但 $n/N > 0.05$ 时，则需要采用有限总体修正系数。除非特别说明，本书中我们假定总体容量相对于样本容量很大，从而无须使用有限总体修正系数。

在第 7.5 节中，我们称 $\bar{x}$ 的标准差为均值的标准误差。一般，标准误差这一术语专指点估计量的标准差。因此，对于比率我们称 $\bar{p}$ 的标准差为比率的标准误差。下面，仍考虑 EAI 公司的例子，计算由 30 名 EAI 管理人员组成的简单随机样本中比率的标准误差。

在 EAI 研究中，我们已知管理人员中参加管理培训计划的总体比率 $p = 0.6$。由于 $n/N = 30/2\,500 = 0.012$，从而在计算 $\bar{p}$ 的标准误差时可以忽略有限总体修正系数。对由 30 名管理人员组成的简单随机样本，有

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.60(1-0.60)}{30}} = 0.089\,4$$

7.6.3 $\bar{p}$ 的抽样分布的形态

现在，我们已经知道了 $\bar{p}$ 的抽样分布的均值和标准差。最后一步是确定抽样分布的形状或形态。样本比率为 $\bar{p} = x/n$。对于一个来自容量很大的总体的简单随机样本而言，样本中具有被关注特征的个体数目 x 是一个服从二项分布的随机变量。由于 n 是一个常数，因此 x/n 的概率等于 x 的二项概率。这就意味着：$\bar{p}$ 的抽样分布也是一个离散型的概率分布，并且 x/n 取每个值的概率等于 x 的概率。

在第 6 章中已经证明，当样本容量足够大并且满足下面两个条件时：

$$np \geqslant 5 \text{ 和 } n(1-p) \geqslant 5$$

二项分布可以用正态分布来近似。假定上述两个条件都满足，则样本比率 $\bar{p} = x/n$ 中 x 的概率分布可以用正态分布来近似。由于 n 是一个常数，$\bar{p}$ 的抽样分布也可以用正态分布来近似。这一近似关系如下：

当 $np \geqslant 5$ 并且 $n(1-p) \geqslant 5$ 时，$\bar{p}$ 的抽样分布可以用正态分布近似。

在实际应用中，当对总体比率进行估计时，我们发现样本容量几乎总是足够大，从而允许对 $\bar{p}$ 的抽样分布进行正态近似。

回忆 EAI 抽样问题中，我们已经知道：参加培训计划的管理人员的总体比率 $p = 0.6$。对于一个容量为 30 的简单随机样本，$np = 30 \times 0.6 = 18$，$n(1-p) = 30 \times 0.4 = 12$。因此，$\bar{p}$ 的抽样分布可以用形如图 7-8 的正态分布近似。

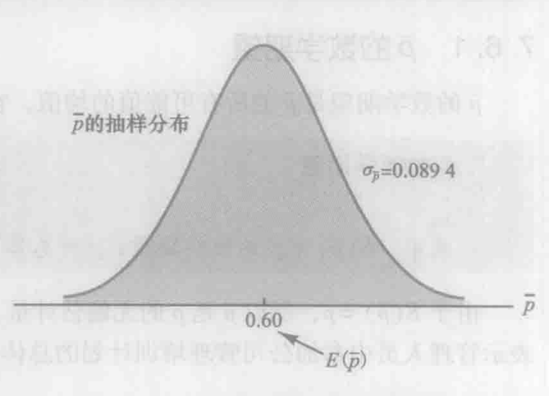


图 7-8 EAI 管理人员中参加管理培训计划的比率 $\bar{p}$ 的抽样分布

7.6.4 $\bar{p}$ 的抽样分布的实际应用

$\bar{p}$ 的抽样分布的应用价值在于，它可以对样本比率与总体比率的差异程度提供概率信息。例如，假定在 EAI 问题中人事部经理需要知道得到的 $\bar{p}$ 值在管理人员参加培训计划的总体比率 p 附近

±0.05 以内的概率是多少？即，样本比率 $\bar{p}$ 介于 0.55～0.65 的概率是多大？图 7-9 中阴影部分的面积即为所求概率。既然 $\bar{p}$ 的抽样分布可用均值为 0.6，标准误差 $\sigma_{\bar{p}}=0.0894$ 的正态分布近似，我们可以在 Excel 中用函数 NORM. DIST 完成这一计算。在 Excel 工作簿的单元格中输入公式“= NORM. DIST（0.65，0.60，0.0894，TRUE）”便可求得与 $\bar{p}=0.65$ 相对应的累积概率值 0.7120。在 Excel 工作簿的单元格中输入公式“= NORM. DIST（0.55，0.60，0.0894，TRUE）”便可求得与 $\bar{p}=0.55$ 相对应的累积概率值 0.2880。因此，样本比率 $\bar{p}$ 在区间（0.55，0.65）的概率是 $0.7120-0.2880=0.4240$。

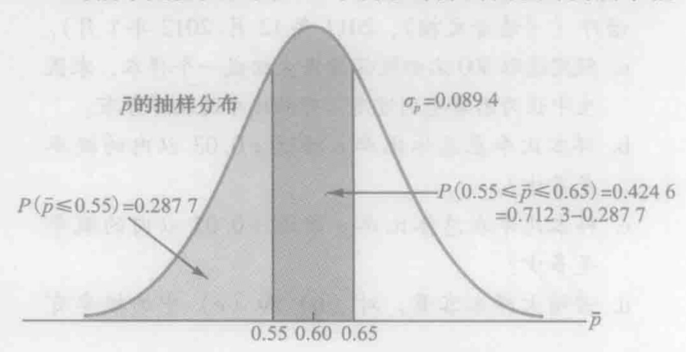


图 7-9 $\bar{p}$ 的值介于 $z=0.55$ 到 $z=0.65$ 之间的概率

我们发现对应于 $\bar{p}=0.65$ 的标准正态随机变量 $z=(0.65-0.6)/0.0894=0.56$。查标准正态概率表，可得与 $z=0.56$ 相对应的累积概率为 0.7123。类似地，当 $\bar{p}=0.55$ 时，$z=(0.55-0.6)/0.0894=-0.56$。查标准正态概率表，得到与 $z=-0.56$ 相对应的累积概率为 0.2877。因此，选取一个样本，其样本比率 $\bar{p}$ 在总体比率 p 附近 0.05 以内的概率是 $0.7123-0.2877=0.4246$。

如果我们考虑将样本容量增加到 $n=100$，则比率的标准误差为

$$\sigma_{\bar{p}}=\sqrt{\frac{0.60(1-0.60)}{100}}=0.0490$$

对由 100 名 EAI 管理人员组成的样本，现在可以计算样本比率的值在总体比率 p 附近 ±0.05 的概率。由于抽样分布近似于均值为 0.60，标准差为 0.0490 的正态分布，我们可以查标准正态概率表得到该概率或面积。当 $\bar{p}=0.65$ 时，$z=(0.65-0.6)/0.049=1.02$，查标准正态概率表，与 $z=1.02$ 相对应的累积概率为 0.8461。类似地，当 $\bar{p}=0.55$ 时，$z=(0.55-0.6)/0.049=-1.02$，与 $z=-1.02$ 相对应的累积概率为 0.1539。因而如果样本容量从 30 增加到 100，那么样本比率 $\bar{p}$ 在总体比率 p 左右 0.05 以内的概率将增加到 $0.8461-0.1539=0.6922$。

练习

方法

☆32. 总体比率为 0.4，从中选取一个样本容量为 200 的样本，用样本比率 $\bar{p}$ 估计总体比率。

a. 样本比率 $\bar{p}$ 落在总体比率 p 附近 ±0.03 以内的概率是多少？

b. 样本比率 $\bar{p}$ 落在总体比率 p 附近 ±0.05 以内的概率是多少？

34. 总体比率为 0.3。对下列样本容量，计算样本比率落在总体比率 p 附近 ±0.04 以内的概率是多少？

a. $n=100$。

b. $n=200$。

c. $n=500$。

d. $n=1000$。

e. 大样本容量的好处是什么？

应用

36. 据《华尔街日报》报道，企业家当中首次创业开办企业时，有 55% 的人的年龄不超过 29 岁，有 45% 的人的年龄大于 30 岁（《华尔街日报》，2012 年 3 月 19 日）。假定选取 200 家企业组成一个样本，了解企业家的重要品质。

a. 求 $\bar{p}$ 的抽样分布，其中 $\bar{p}$ 为企业家首次创业年龄不超过 29 岁的样本比率。

b. （a）中的样本比率在总体比率 p 附近 ±0.05 以内的概率是多少？

c. 假定选取 200 家企业组成一个样本，了解企业家的重要品质。求 $\bar{p}$ 的抽样分布，其中 $\bar{p}$ 为企业家首次创业年龄达到或超过 30 岁的样本比率。

d. （c）中的样本比率在总体比率 p 附近 ±0.05 以内的概率是多少？

e. （b）和（d）中的概率有不同吗？如果不同，为什么？

f. 当样本容量为 400 时，回答（b）中问题。概率更小吗？为什么？

38. 有42%的初级保健医生认为患者接受了不必要的治疗（《读者文摘》，2011年12月/2012年1月）。
    a. 假定选取300名初级保健医生组成一个样本，求医生中认为患者受到过度治疗的比率的抽样分布。
    b. 样本比率在总体比率 p 附近 ± 0.03 以内的概率是多少？
    c. 样本比率在总体比率 p 附近 ± 0.05 以内的概率是多少？
    d. 若增大样本容量，对（b）和（c）中的概率有何影响？为什么？
40. 据美国百货制造商报道，76%的顾客阅读产品标签所列示的配方。假定总体比率 $p=0.76$，从总体中选取400名顾客组成一个样本。
    a. 求 $\bar{p}$ 的抽样分布。其中 $\bar{p}$ 是样本中阅读产品标签所列示配方的顾客所占的比率。
    b. 样本比率落在总体比率 p 附近 ± 0.03 以内的概率是多少？
    c. 当样本是由750名顾客组成时，回答（b）中问题。

7.7 点估计的性质

本章我们已经说明了样本统计量，如样本均值 $\bar{x}$，样本标准差 s，样本比率 $\bar{p}$ 如何用做相应总体参数 μ，σ 和 p 的点估计量。直观上，用这些样本统计量做为相应总体参数的点估计量是很有吸引力的。然而，在一个样本统计量做为点估计量之前，统计学家需要检查该样本统计量是否具有好的点估计量应具备的性质。本节我们讨论一个好的点估计量应该具有的三条性质：无偏性、有效性和一致性。

由于有许多不同的样本统计量可用做总体不同参数的点估计量，因此本节中我们采用如下这种一般的记号。

θ——感兴趣的总体参数

$\hat{\theta}$——样本统计量或 θ 的点估计量

符号 θ 是希腊字母，读做 theta；符号 $\hat{\theta}$ 读做 theta 尖。通常，θ 代表任一总体的参数，比如总体均值、总体标准差和总体比率等；$\hat{\theta}$ 代表相应的样本统计量，比如样本均值、样本标准差和样本比率。

7.7.1 无偏性

如果样本统计量的期望值等于所估计的总体参数，则称该样本统计量是相应总体参数的无偏估计量。

无偏性

样本统计量 $\hat{\theta}$ 是总体参数 θ 的无偏估计量，如果

$$E(\hat{\theta}) = \theta$$

式中，$E(\hat{\theta})$ 代表样本统计量 $\hat{\theta}$ 的数学期望。

于是，样本无偏统计量的所有可能值的数学期望或均值等于被估计的总体参数。

图7-10展示了有偏和无偏点估计量。在无偏估计量的图示中，抽样分布均值与总体参数的值相等。此时，由于有时点估计量大于 θ，有时小于 θ，因此估计的误差相抵。在有偏估计量的情形下，抽样分布的均值大于或者小于总体参数的值。在图7-10b中，由于 $E(\hat{\theta})$ 比 θ 大，从而样本统计量以较大的概率高估总体参数值。偏斜程度如图所示。

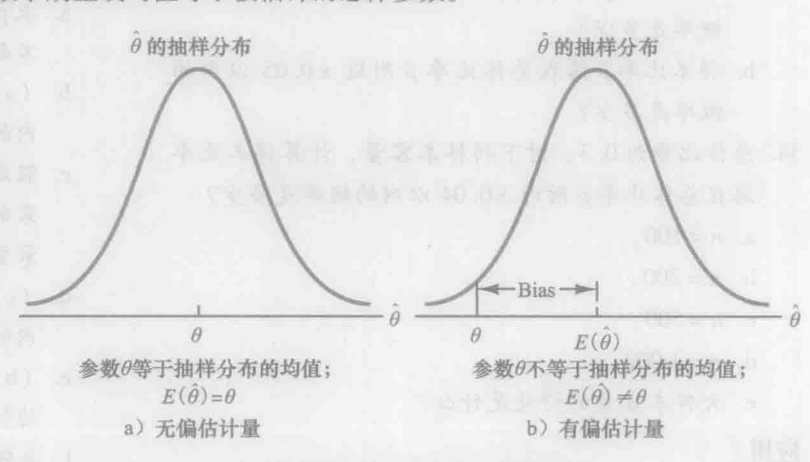


图7-10 有偏与无偏点估计量的例子

在讨论样本均值和样本比率的抽样分布时，我们证明了 $E(\bar{x})=\mu$ 和 $E(\bar{p})=p$。因此 $\bar{x}$ 和 $\bar{p}$ 都是相应总体参数 μ 和 p 的无偏

估计量。

对于样本标准差 s 和样本方差 s^2，可以证明 $E(s^2)=\sigma^2$，因此样本方差 s^2 是总体方差 σ^2 的无偏估计量。实际上，当我们在第3章中第一次给出样本方差和样本标准差的公式时，分母是 $n-1$ 而不是 n，用 $n-1$ 而不用 n 的原因正是为了使样本方差是总体方差的无偏估计量。

7.7.2 有效性

假定一个简单随机样本由 n 个个体组成，给出了总体同一参数的两个不同的无偏点估计量。这时，我们倾向于采用标准误差较小的点估计量，因为它给出的估计值与总体参数更接近。称有较小标准误差的点估计量比其他点估计量有更相对有效（relative efficiency）。

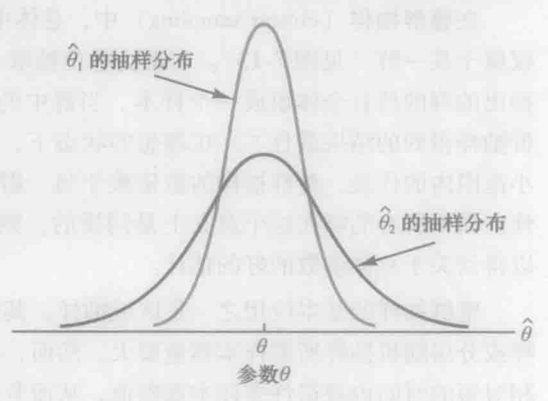


图7-11 两个无偏点估计量的抽样分布

图7-11 给出了两个无偏点估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 的抽样分布。注意，由于 $\hat{\theta}_1$ 的标准误差比 $\hat{\theta}_2$ 的标准误差小，因此 $\hat{\theta}_1$ 的值比 $\hat{\theta}_2$ 的值接近参数 θ 的机会更大。[㊀] 由于点估计量 $\hat{\theta}_1$ 的标准误差比点估计量 $\hat{\theta}_2$ 的标准误差小，因此 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 相对更有效，是更好的点估计量。

7.7.3 一致性

一个好的点估计应该具备的第三条性质是**一致性**（consistency）。粗略地讲，如果随着样本容量的增大，点估计量的值与总体参数越来越接近，则称该点估计量是一致的。换而言之，大样本情形下比小样本情形下更易于得到一个好的点估计。注意，对样本均值 $\bar{x}$，我们可以证明 $\bar{x}$ 的标准误差 $\sigma_{\bar{x}}=\sigma/\sqrt{n}$。由于 $\sigma_{\bar{x}}$ 与样本容量相关，样本容量越大 $\sigma_{\bar{x}}$ 值越小，因此我们得出结论，大样本容量下所给出的点估计与总体均值 μ 更接近。从这个意义上，我们可以说样本均值 $\bar{x}$ 是总体均值 μ 的一个一致估计量。同理，我们也可以得出结论，样本比率 $\bar{p}$ 是总体比率 p 的一个一致估计量。

注释和评论

在第3章，我们证明了均值和中位数是度量位置中心的两种方法。本章，我们只讨论了均值，这是因为在正态总体中总体均值与总体中位数是相等的，当从正态总体中抽样时，中位数的标准误差大约比均值的标准误差大将近25%。回忆在EAI问题中，$n=30$，均值的标准误差 $\sigma_{\bar{x}}=730.3$，此时中位数的标准误差大约为 $1.25\times730.30=913$。因此，样本均值更有效，以更高的概率落入总体均值附近的某一特定范围。

7.8 其他抽样方法

作为自有限总体进行抽样的一种方法，我们已经描述了简单随机抽样并且讨论了简单随机抽样中 $\bar{x}$ 和 $\bar{p}$ 的抽样分布的性质。除此之外，还有分层随机抽样、整群抽样和系统抽样等方法。在某些情况下，它们优于简单随机抽样。本节我们简要介绍这些其他的抽样方法。详细讨论见本教材所附光盘中的第22章。[㊁]

7.8.1 分层随机抽样

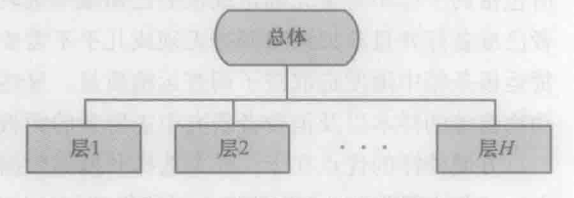


图7-12 分层随机抽样的图示

在**分层随机抽样**中，总体中的个体首先被分成称作层的组，总体中的每一个体属于且仅属于某一层。由抽样设计者自行进行层的划分，比如：按照部门、位置、年龄、产业类型等进行划分。当每一层内的个体都尽可能地相似时，得到的结果最佳。图7-12是将一

㊀ 当自正态总体抽样时，样本均值的标准误差比样本中位数的标准误差要小，因此样本均值比样本中位数有效。

㊁ 本节对除简单随机抽样以外的其他抽样方法做一个简介。

个总体分成 H 层的图示。

分层以后，从每一层抽一个简单随机样本。将每层的样本的结果合并起来，利用公式对感兴趣的总体参数进行估计。分层随机抽样的值依赖于层内个体的同质性。如果层内的个体是同质的，该层有较低的方差，那么在相对小的样本容量下便可获得层特征的一个好的估计[㊀]。如果各层是同质的，那么分层随机抽样方法能在较小的样本容量下得到与简单随机抽样同样精确的结果。

7.8.2 整群抽样

在**整群抽样**（cluster sampling）中，总体中的个体首先被分成称作群的单个组，总体中的每一个个体属于且仅属于某一群（见图 7-13）。以群为单位抽取一个简单随机样本，抽出的群的所有个体组成一个样本。当群中的个体不同质时，整群抽样得到的结果最佳[㊁]。在理想的状态下，每一群是整个总体小范围内的代表。整群抽样的值依赖于每一群对整个总体的代表性。如果所有的群在这个意义上是同质的，则抽取小量的群就可以得到关于总体参数的好的估计。

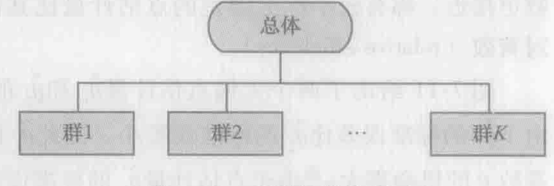


图 7-13 整群抽样的图示

整群抽样的基本应用之一是区域抽样，其中群为街区或以其他方式定义的区域。整群抽样通常比简单随机抽样或分层随机抽样所需样本容量要大。然而，事实上当派一个采访员去一个样本群（如，城市路口）时，可以在相对短的时间内获得许多样本观察值，从而节约费用。因此，可以在更低的总成本下获得更大量的样本。

7.8.3 系统抽样

在某些抽样情况下，特别是对那些容量很大的总体时，如果采取先找到一个随机数再计数或在总体个体的清单中查找相应的个体来选取一个简单随机样本的话，这是非常费时的。简单随机抽样的另一种替代方法为**系统抽样**（systematic sampling）。例如，如果希望从含 5 000 个个体的总体中选取样本容量为 50 的样本，我们从总体每 5 000/50 = 100 个个体中选出一个为样本点。在系统抽样情形下，即为在总体清单的前 100 个个体中随机选取一个，然后从第一个已选出的个体开始，依次向下，在总体清单中每隔 100 个个体选取一个为样本点。实际上，从第一个选取的个体开始向后每隔 100 个个体选取一个做为样本点，通过机械地在总体中移动得到 50 个样本点。通常，采用这种方法比用简单随机抽样方法来选取 50 个样本点要容易些。由于第一个被选中的个体是随机的，通常假定系统抽样具有简单随机抽样的性质。当总体中个体的排列是按照个体的随机顺序排列时，这一假设尤其适用。

7.8.4 方便抽样

以上讨论的抽样方式指的均属于概率抽样技术，从总体中选出的个体以已知的概率入选样本。概率抽样的优点在于样本统计量的抽样分布通常是已知的。本章所给出的简单随机抽样的公式可用于确定抽样分布的性质。于是，对基于样本结果推断总体时所产生的误差，可利用抽样分布作出概率解释。

方便抽样（convenience sampling）是一种非概率抽样技术。顾名思义，样本的确定主要是基于简便。样本中所包括的个体不是事先确定或按照已知概率选取的。例如，一名教授在某所大学做一项调查，由于学生中的志愿者已准备好并且参加该项调查无须或几乎不需要成本，故选择由他们组成样本。类似地，一个监督员可以从许多货运板条箱中随便选取橙子调查运输质量。显然，给每个橙子标签，用概率抽样方式是不现实的。再比如，野生动物营地的样本以及消费者研究中志愿者的面板数据都是方便样本。

方便抽样的优点在于，样本选择和搜集数据相对容易。然而，从其对总体的代表性来讲，它不能用于估计样本的“拟合性”。一个方便样本可能得到好的结论，也可能不能。没有统计上公认的方法可用于对抽样结果的质

㊀ 当每层当中个体的方差相当小时，分层随机抽样的效果最好。
㊁ 当每群是总体一个小范围的代表时，整群随机抽样的效果最好。

量进行概率分析和推断。有时，研究者将概率抽样中所设计的统计方法用于方便抽样，认为可以将方便抽样视为概率抽样。然而，这种观点尚未得到认可，我们在用方便抽样的结果对总体进行推断时，在解释上要谨慎。

7.8.5 判断抽样

另一种非概率抽样技术为**判断抽样**（judgment sampling）。在这种抽样方法中，由对研究总体非常了解的人主观确定选择总体中他认为最具代表性的个体组成样本。通常，这种方法在选取样本时相对容易。例如，报告者可抽样两个或三个议员，认为这些议员的想法反映了全体议员的普遍意见。然而，抽样结果的质量依赖于选择样本的人的判断。同样，基于判断抽样对总体进行推断时，在做结论的时候要特别小心。

注释和评论

当自有限总体进行抽样时，我们推荐采用下列概率抽样方法：简单随机抽样、分层随机抽样、整群抽样或系统抽样。对这些方法，有公式可用于评判抽样结果与总体特征的接近程度——“拟合度”。对方便抽样或者判断抽样而言，则不能评判抽样结果的“拟合度”。因而，对于由非概率抽样方法得到的结果，在解释上要特别小心。

小结

本章，我们给出了抽样和抽样分布的概念。我们说明了如何从有限总体抽取一个简单随机样本和如何从无限总体抽取一个随机样本，利用收集的样本数据可以对总体参数进行点估计。由于不同的样本给出了不同的点估计量的值，因此点估计量，比如 $\bar{x}$ 和 $\bar{p}$，都是随机变量。这些随机变量的概率分布叫作抽样分布。我们特别描述了样本均值 $\bar{x}$ 和样本比率 $\bar{p}$ 的抽样分布。

在考虑 $\bar{x}$ 和 $\bar{p}$ 的抽样分布的特征时，我们证明了 $E(\bar{x})=\mu$ 和 $E(\bar{p})=p$。因此，$\bar{x}$ 和 $\bar{p}$ 是无偏估计量。在给出估计量的标准差或标准误差的公式后，我们给出了 $\bar{x}$ 和 $\bar{p}$ 的抽样分布服从正态分布的必要条件。其他抽样方式还包括分层随机抽样、整群抽样、系统抽样、方便抽样和判断抽样，对此我们也进行了讨论。

关键术语

sampled population 抽样总体 抽取样本的总体。

frame 抽样框 抽取样本时所用的个体清单。

parameter 参数 总体的数值特征，如总体均值 μ、总体标准差 σ、总体比率 p，……

simple random sample 简单随机样本 从容量为 N 的有限总体中抽取容量为 n 的一个样本，使得样本中的每一个个体都以相同的概率被抽到。

sampling without replacement 无放回抽样 一个个体一旦被选入样本，就从总体中剔除，不能再次被选入样本。

sampling with replacement 有放回抽样 一个个体被选入样本后仍然放回总体中。先前被抽到的个体可能再次被选入样本，从而在样本中多次出现。

random sample 随机样本 如果从无限总体中抽取一个容量为 n 的样本，使得下面的条件得到满足：1. 抽取的每个个体来自同一总体；2. 每个个体的抽取是独立的。则称该样本是一个随机样本。

sample statistic 样本统计量 一种样本特征，如样本均值 $\bar{x}$、样本标准差 s、样本比率 $\bar{p}$，等等。样本统计量的值用于估计相应总体参数。

point estimator 点估计量 提供总体参数点估计的样本统计量，如 $\bar{x}$、s 或 $\bar{p}$。

point estimate 点估计值 点估计量的值，在一个特定实例中用来作为总体参数的估计值。

target population 目标总体 进行统计推断（比如：点估计）的总体。目标总体应该与相应的抽样总体尽可能多的相似，这在统计推断中尤为重要。

sampling distribution 抽样分布 一个样本统计量所有可能值构成的概率分布。

unbiased 无偏性 点估计量的一个性质，此时点估计量的数学期望等于所估总体参数的值。

finite population correction factor 有限总体修正系数 当从有限总体而非无限总体抽样时，$\sigma_{\bar{x}}$ 和 $\sigma_{\bar{p}}$ 的公式中出现的项 $\sqrt{(N-n)/(N-1)}$。根据经验法则，当 $n/N \leq 0.05$ 时一般可以忽略有限总体修正系数。

standard error 标准误差 点估计量的标准差。

central limit theorem 中心极限定理 一个定理，当样本容量很大的时候可以用正态概率分布近似 $\bar{x}$ 的抽样分布。

relative efficiency 相对有效性 对同一总体参数的两个无偏的点估计量，称有更小标准差的点估计量更有效。

consistency 一致性 点估计量的一个性质，随着样本容量的增大，点估计值与总体参数越来越接近。

stratified random sampling 分层随机抽样 一种概率抽样方式，先将总体分成若干层，然后在每层中进行简单随机抽样。

cluster sampling 整群抽样 一种概率抽样方式，先将总体分成若干群，然后以群为单位进行简单随机抽样。

systematic sampling 系统抽样 一种概率抽样方式，从头 k 个个体中随机选一个，然后往后每隔 k 个个体选取一个个体进入样本。

convenience sampling 方便抽样 一种非概率抽样方式，基于简便选择个体组成样本。

judgment sampling 判断抽样 一种非概率抽样方式，基于研究人员的判断选择个体组成样本。

重要公式

$\bar{x}$ 的数学期望

$$E(\bar{x}) = \mu \tag{7-1}$$

$\bar{x}$ 的标准差（标准误差）

有限总体 无限总体

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{7-2}$$

$\bar{p}$ 的数学期望

$$E(\bar{p}) = p \tag{7-4}$$

$\bar{p}$ 的标准差（标准误差）

有限总体 无限总体

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{7-5}$$

补充练习

42. 《美国新闻和世界报道》公布了美国最好的大学的详细信息（《美国最好的大学》，2009 年）。除此之外，他们还列出了全美最好的 133 所国立大学的名单。你想要从这些大学中选取一个样本，并对他们的学生进行后续研究。从表 7-1 中第三列末端的随机数开始，忽略五位数中的前两位只利用三位数，从 959 开始按列向上从编号 1～133 的大学中选取出 7 所大学组成一个简单随机样本。有必要的话，可继续沿着第四列和第五列末端随机数开始向上进行选取。

44. Foot Lock 用每平方英尺年销售额来衡量商铺的绩效。目前，每平方英尺的年销售额达到 406 美元（《华尔街日报》，2012 年 3 月 7 日）。管理层要求你对 64 家商铺组成的一个样本进行研究。Foot Lock 共有 3 400 家商铺，假设商铺每平方英尺年销售额的总体标准差为 80 美元。
    a. 选取 Foot Lock 的 64 家商铺组成一个样本，试求样本均值 $\bar{x}$ 的抽样分布，其中 $\bar{x}$ 是每平方英尺年销售额的样本均值。
    b. 样本均值在总体均值附近 ±15 美元以内的概率为多少？
    c. 假定你得到的样本均值为 380 美元。求样本均值不高于 380 美元的概率。你认为这个样本是由低业绩的异常店铺组成的吗？

46. 扣除必要的补助金后，南加州大学（USC）的平均费用为 27 175 美元（《美国新闻和世界报道》，美国最好的大学，2009 年）。假定总体标准差为 7 400 美元，从总体中抽取 60 名 USU 学生组成一个简单随机样本。
    a. 均值的标准误差是多少？
    b. 样本均值大于 27 175 美元的概率是多少？
    c. 样本均值在总体均值 $\bar{x}$ 附近 ±1 000 美元以内的概率为多少？
    d. 若样本容量增大到 100，则（c）中概率为多少？

48. 某研究人员调查结果表明，均值的标准误差为 20，而总体的标准差为 500。试问：
    a. 调查中的样本容量为多大？
    b. 估计量落在总体均值 μ 左右 ±25 以内的概率为多大？

50. 私人公司中大约有 28% 为女性所拥有（*The Cincinnati Enquirer*，2006 年 1 月 26 日）。根据 240 家私

人公司组成的一个样本，回答下列问题。

a. 试求 $\bar{p}$ 的抽样分布，其中 $\bar{p}$ 是女性所拥有企业的样本比率。

b. 样本比率 $\bar{p}$ 在总体比率 p 附近 ±0.04 以内的概率是多少？

c. 样本比率 $\bar{p}$ 在总体比率 p 附近 ±0.02 以内的概率是多少？

52. 想要登广告的人与互联网服务商以及搜索引擎签订合同后将广告放在网站上，并根据点击其广告的潜在顾客的数目支付费用。不幸的是，仅仅为了提高广告收入而发生的欺诈性点击行为已经成为一个严重的问题。登广告的人中有40%认为他们是欺诈性点击的受害者(《商业周刊》，2006 年 3 月 13 日)。假定选取380 名登广告的人组成一个简单随机样本，以便更多地了解欺诈性点击对他们的影响。求：

a. 欺诈性点击的样本比率在总体比率附近 ±0.04 之内的概率。

b. 欺诈性点击的样本比率大于0.45 的概率。

54. Lori Jeffrey 是大学教材主要出版商的一名销售代表，她的工作很出色。历史数据显示，在 Lori 的电话销售中教材被采用的可能性达到 25%。观察她一个月的电话销售记录，并将销售电话的结果组成一个样本。假设数据的统计分析显示，比率的标准误差为 0.062 5。

a. 在这项分析中，样本容量为多大？也就是说，在一个月里 Lori 打了多少个销售电话？

b. 令 $\bar{p}$ 代表在这个月期间 Lori 电话销售中教材被采用的样本比率，求 $\bar{p}$ 的抽样分布。

c. 利用 $\bar{p}$ 的抽样分布，求在一个月内 Lori 电话销售中教材被采用的比率不低于 30% 的概率。

第8章 区间估计

CHAPTER 8

实践中的统计

Food Lion ㊀

北卡罗来纳州，索尔兹伯里

Food Lion 食品城创立于 1957 年，是美国最大的连锁超市之一。在美国东南部和大西洋中部地区的 11 个州中有 1 300 家连锁店。公司销售的产品超过 24 000 种，提供在全美和地区进行广告宣传的名牌商品，其中由 Food Lion 特别供给的具有自主商标的高品质产品在数量上逐年增加。Food Lion 通过高效的运作，比如标准化的存储方式、仓库的创新设计、高效能的设备以及与供应商的同步数据，来确保物美价廉。Food Lion 承诺，未来将一如既往地保持持续的创新、发展、价格领先和对顾客的优质服务。

在这样一个库存密集型的企业中，公司决定采取后进先出（LIFO）的库存计价法。该方法将当期成本与当期收益相配比，从根本上降低了价格剧烈的变化对损益的影响。另外，在通货膨胀时期，LIFO 库存计价方法可以减少净收益，从而减少所得税。

Food Lion 有七座存货库房，分别存放杂货、纸/家居、宠物、保健和化妆品、奶制品、烟草和酒类。Food Lion 为每个库房中建立了 LIFO 指数。例如，杂货存货库房的 LIFO 指数为 1.008，说明最近一年当中由于通货膨胀的原因，致使公司以当期成本计价的杂货的库存价值增加了 0.8%。

对每个存货库房，LIFO 指数要求对每种产品的年末存货按当年年末成本和上一年年末成本分别计价。为避免在全部的 1 200 个存储点中对存货逐一进行计数而浪费财力和时间，Food Lion 选取 50 个存储点组成一个随机样本。年末仅对这些样本存储点中的库存进行实地盘存。每个存储点根据每种产品的当年成本和上一年成本构造 LIFO 指数。

去年，保健和化妆品存货库房的 LIFO 指数的样本估计值为 1.015。在 95% 的置信水平下，样本估计中 Food Lion 计算得到的边际误差为 0.006。于是，总体 LIFO 指数的 95% 置信水平的区间估计为（1.009，1.021）。可以证明这一精度是很不错的。

本章你将学习如何计算与样本估计相联系的边际误差，以及如何使用这些信息构造并解释总体均值和总体比率的区间估计。

在第 7 章，我们发现点估计量是用于估计总体参数的样本统计量。例如，样本均值 $\bar{x}$ 是总体均值 μ 的点估计量，样本比率 $\bar{p}$ 是总体比率 p 的点估计量。因为我们不可能期望点估计量能给出总体参数的精确值，所以经常在点估计上加减一个被称为**边际误差**（marginal of error）的值来计算**区间估计**（interval estimate）。区间估计的一般形式如下：

$$\text{点估计} \pm \text{边际误差}$$

区间估计的目的在于，提供基于样本得出的点估计值与总体参数值的接近程度方面的信息。

本章，我们将说明如何对总体均值 μ 和总体比率 p 进行区间估计。总体均值的区间估计的一般形式为

$$\bar{x} \pm \text{边际误差}$$

类似地，总体比率的区间估计的一般形式为

$$\bar{p} \pm \text{边际误差}$$

在计算这些区间估计时，$\bar{x}$ 和 $\bar{p}$ 的抽样分布起着非常重要的作用。

8.1 总体均值的区间估计：σ 已知情形

为了对总体均值进行区间估计，必须利用总体标准差 σ 或者样本标准差 s 计算边际误差。在大多数的应用中 σ 是未知的，于是 s 用于计算边际误差。但是在一些应用中，我们在抽样前可以根据大量有关的历史数据估计总

㊀ 作者感谢 Food Lion 公司的税务主管 Keith Cunningham 和税务会计 Bobby Harkey，他们为“实践中的统计”提供了本案例。

体标准差。比如，在质量控制应用中，若假定生产过程是正常运行或者“控制之中”，则可以看作总体标准差是已知的。我们称这种情形为 σ 已知（σ known）。本节中，我们举例说明在 σ 已知情形下如何构造区间估计。

Lloyd 百货公司每周选择 100 名顾客组成一个简单随机样本，目的在于了解他们每次购物的消费额。令 x 表示每次购物的消费额，样本均值 $\bar{x}$ 是 Lloyd 全体顾客每次购物消费额的总体均值 μ 的点估计。Lloyd 公司的这项周度调查已经进行了许多年。根据历史数据，Lloyd 假定总体标准差已知，为 $\sigma=20$ 美元。并且历史数据还显示总体服从正态分布。

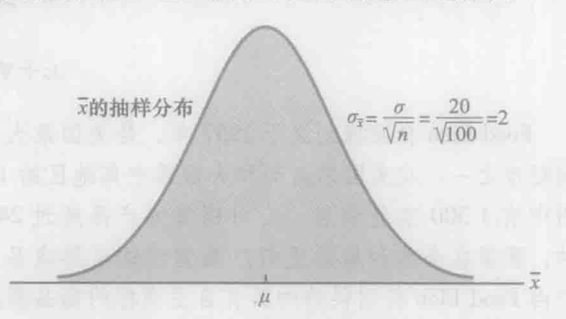


图 8-1　100 名顾客的简单随机样本中，消费额的样本均值的抽样分布

最近一周，Lloyd 调查了 100 名顾客（$n=100$），得到样本均值 $\bar{x}=82$ 美元。每次购物消费额的样本均值是总体均值 μ 的点估计。在下面的讨论中，我们将介绍如何计算估计的边际误差，以及如何建立总体均值的区间估计。

8.1.1　边际误差和区间估计

在第 7 章中，我们发现可以利用 $\bar{x}$ 的抽样分布计算 $\bar{x}$ 在 μ 附近一定范围内的概率。在 Lloyd 公司的例子里，历史数据表明消费额总体服从标准差 $\sigma=20$ 的正态分布。因此，利用第 7 章的知识可知，$\bar{x}$ 的抽样分布服从标准误差 $\sigma_{\bar{x}}=\sigma/\sqrt{n}=20/\sqrt{100}=2$ 的正态分布，该抽样分布如图 8-1 所示[⊖]。因为抽样分布说明 $\bar{x}$ 的值如何分布在总体均值 μ 附近，所以 $\bar{x}$ 的抽样分布提供了关于 $\bar{x}$ 与 μ 之间可能存在的差别的信息。

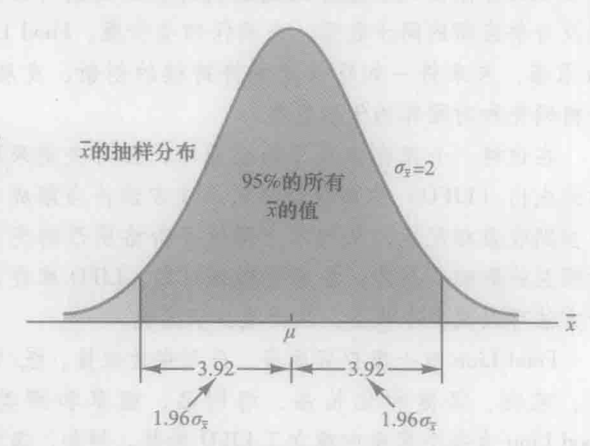


图 8-2　$\bar{x}$ 的抽样分布：样本均值在 $\mu\pm3.92$ 以内的区域

查标准正态概率表后我们发现，任何正态分布随机变量都有 95% 的值在均值附近 ±1.96 个标准差以内。因此，当 $\bar{x}$ 的抽样分布是正态分布时，一定有 95% 的 $\bar{x}$ 的值在均值 $\mu\pm1.96\sigma_{\bar{x}}$ 以内。在 Lloyd 公司的例子里，我们已知 $\bar{x}$ 的抽样分布是正态分布并且标准误差 $\sigma_{\bar{x}}=2$。因为 $1.96\sigma_{\bar{x}}=1.96\times2=3.92$，所以在 $n=100$ 的样本容量下，$\bar{x}$ 的所有值中有 95% 落在总体均值 μ 附近 ±3.92 以内（见图 8-2）。

在前文中，总体均值 μ 的区间估计的一般形式为 $\bar{x}\pm$ 边际误差。在 Lloyd 公司的例子里，假定令边际误差等于 3.92，利用 $\bar{x}\pm3.92$ 计算 μ 的区间估计。为了解释这一区间估计，我们抽取三个不同的简单随机样本，每一个样本由 100 名 Lloyd 公司的顾客组成，考虑得到的三个 $\bar{x}$ 值。第一个样本的样本均值如图 8-3 所示。在这种情形下，图 8-3 表明 $\bar{x}_1\pm3.92$ 构建的区间包含了总体均值 μ。如果第二个样本均值 $\bar{x}_2$ 如图 8-3 所示，现在考虑将会发生什么。虽然样本均值与第一个样本均值不同，但我们看到 $\bar{x}_2\pm3.92$ 构建的区间仍包括总体均值 μ。但是，如果第三个样本均值 $\bar{x}_3$ 如图 8-3 所示，则这时又会怎样呢？在这种情形下，$\bar{x}_3\pm3.92$ 构建的区间不

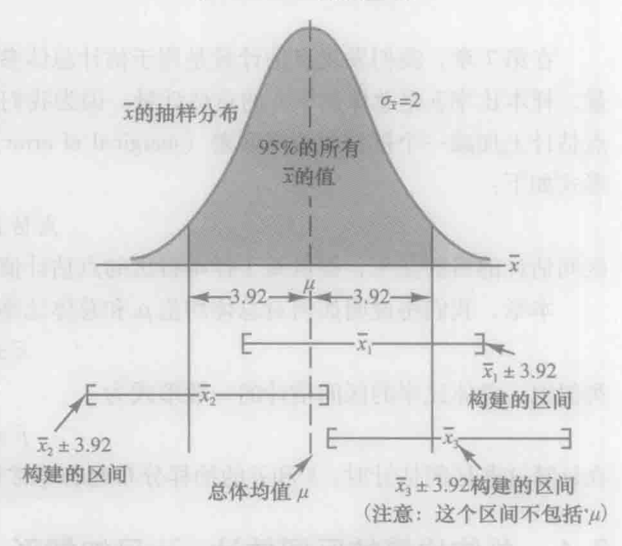


图 8-3　抽取的样本均值分别为 $\bar{x}_1$、$\bar{x}_2$ 和 $\bar{x}_3$ 时所构造的区间

⊖ 我们利用消费额总体服从正态分布这一事实，得出 $\bar{x}$ 的抽样分布服从正态分布的结论。如果总体不是服从正态分布，则我们可以根据中心极限定理及样本容量 $n=100$ 得出 $\bar{x}$ 的抽样分布近似服从正态分布。图 8-1 给出了在这两种情形下 $\bar{x}$ 的抽样分布。

包括总体均值μ。此时，由于$\bar{x}_3$落在抽样分布的上侧，偏离μ的距离超过3.92，所以$\bar{x}_3 \pm 3.92$构建的区间不包含总体均值μ。

在图8-3中，阴影区域内的任一样本均值$\bar{x}$所构造的区间均包含总体均值μ。由于所有样本均值中有95%落在此阴影区域中，所以$\bar{x} \pm 3.92$所构建的所有区间中有95%的区间包含总体均值μ。

在最近一周，Lloyd的质量保证部门对100名顾客进行了调查，得到消费额的样本均值$\bar{x}=82$。利用$\bar{x} \pm 3.92$构造区间估计，我们得到82 ± 3.92。因此，根据最近一周的数据，μ的区间估计为$82-3.92=78.08$到$82+3.92=85.92$。因为，利用$\bar{x} \pm 3.92$构造的所有区间中有95%包括总体均值，所以我们说：有95%的把握相信区间（78.08，85.92）包括总体均值μ。我们称这个区间是在95%**置信水平**（confidence level）下建立的，其中数值0.95称作**置信系数**（confidence coefficient），区间（78.08，85.92）称作95%**置信区间**（confidence interval）[⊖]。

边际误差由$z_{\alpha/2}(\sigma/\sqrt{n})$给出，下面是$\sigma$已知情形下总体均值区间估计的一般形式。

总体均值的区间估计：σ已知

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8-1}$$

式中，$1-\alpha$为置信系数；$z_{\alpha/2}$表示标准正态概率分布上侧面积为$\alpha/2$时的z值。

在Lloyd公司例子中，我们利用式（8-1）构造95%的置信区间。对于一个95%的置信区间，置信系数$1-\alpha=0.95$。于是，$\alpha=0.05$。查标准正态概率表，上侧面积为$\alpha/2=0.05/2=0.025$时对应的$z_{0.025}=1.96$。Lloyd公司的样本均值$\bar{x}=82$，$\sigma=20$，样本容量$n=100$。我们得到

$$82 \pm 1.96\frac{20}{\sqrt{100}} \qquad 82 \pm 3.92$$

于是，利用式（8-1），当边际误差为3.92时，95%的置信区间为从$82-3.92=78.08$到$82+3.92=85.92$。

虽然95%是经常使用的置信水平，但也可以考虑采用其他的置信水平，比如90%和99%。表8-1中给出了最常用置信水平下的$z_{\alpha/2}$值。利用这些值和式（8-1），对于Lloyd问题，公司例子中90%的置信区间为

$$82 \pm 1.645\frac{20}{\sqrt{100}} \qquad 82 \pm 3.29$$

表8-1 最常用的置信水平下的$z_{\alpha/2}$值

置信水平	α	$\alpha/2$	$z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.960
99%	0.01	0.005	2.576

因此，在90%的置信水平下，边际误差为3.29，置信区间为从$82-3.29=78.71$到$82+3.29=85.29$。类似地，99%的置信区间为

$$82 \pm 2.576\frac{20}{\sqrt{100}} \qquad 82 \pm 5.15$$

于是，在99%的置信水平下，边际误差为5.15，置信区间为从$82-5.15=76.85$到$82+5.15=87.15$。

比较90%、95%和99%的置信水平的结果，我们看到：要想达到较高的置信水平，必须加大边际误差，即加大置信区间的宽度。

8.1.2 应用中的建议

如果总体服从正态分布，式（8-1）所给出的置信区间是精确的。换言之，如果利用式（8-1）反复计算95%的置信区间，则得到的置信区间中恰好有95%的区间包含总体均值。如果总体不服从正态分布，则式（8-1）所给出的置信区间是近似的。在这种情形下，近似的程度依赖于总体的分布和样本容量。

⊖ 讨论说明了区间叫作95%的置信区间的理由。

在绝大部分应用中，当利用式（8-1）建立总体均值的区间估计时，样本容量 $n \geqslant 30$ 已经足够。然而，如果总体的分布不服从正态分布但是大致对称，则样本容量至少为 15 时才能得到置信区间一个好的近似。仅当分析人员坚信或者愿意假设总体分布至少是近似正态时，才可以在更小的样本容量下利用式（8-1）。

注释和评论

1. 本节所讨论的区间估计方法是在总体标准差 σ 已知的假设下进行的。σ 已知意味着，在估计总体均值时，我们进行抽样之前可以利用历史数据或者其他信息得到总体标准差 σ 的一个好的估计。所以，从技术上讲，我们并不能认为 σ 在实际中是确定已知的。这只是意味着，我们在进行抽样之前得到了总体标准差的一个好的估计，于是，不必利用同一样本同时估计总体均值和总体标准差。
2. 在区间估计的表达式（8-1）中，样本容量 n 出现在分母上。于是，实际应用中当某一样本容量产生的区间太宽时，我们可以考虑增大样本容量。由于 n 出现在分母，增大样本容量可以使边际误差减小，使区间变窄，精度提高。为了达到所希望的精度，如何确定简单随机样本容量的方法将在第 8.3 节中讨论。

练习

方法

☆2. 一个简单随机样本由 50 项组成，样本均值 $\bar{x}=32$，总体标准差 $\sigma=6$。

a. 求总体均值的 90% 的置信区间。

b. 求总体均值的 95% 的置信区间。

c. 求总体均值的 99% 的置信区间。

4. 已知总体均值的 95% 置信区间为（152，160）。如果 $\sigma=15$，则研究中应选用多大的样本容量？

应用

6. 尼尔森媒体调查对在晚间 8:00 ~ 11:00 家庭看电视的时间进行了研究，研究中所用的数据存放在名为 Nielsen 的文件中（《世界年鉴》，2003 年）。根据过去的研究，假定总体标准差已知，为 $\sigma=3.5$ 个小时。试求每个家庭在晚间 8:00 ~ 11:00，每周看电视时间均值的 95% 的置信区间估计。

8. 研究表明，按摩疗法对健康保健益处颇多并且不是太贵（《华尔街日报》，2012 年 3 月 13 日）。10 次一小时按摩费用组成一个样本，平均费用为 59 美元。一小时按摩费用的总体标准差为 $\sigma=5.50$ 美元。

a. 如果想要计算边际误差，则需要对总体作出怎样的假设。

b. 在 95% 的置信水平下，边际误差是多少？

c. 在 99% 的置信水平下，边际误差是多少？

10. 据 *Playbill* 杂志报道，其读者中家庭年均收入为 119 115 美元（*Playbill*，2006 年 1 月）。假设家庭年收入的平均值是基于 80 个家庭组成的样本估计得到的，并且根据以往的研究可知总体标准差 $\sigma=$ 30 000 美元。

a. 求总体均值的 90% 置信区间估计。

b. 求总体均值的 95% 置信区间估计。

c. 求总体均值的 99% 置信区间估计。

d. 当置信水平增大时，置信区间的宽度如何变化？这合理吗？为什么？

8.2 总体均值的区间估计：σ 未知情形

在建立总体均值的区间估计时，我们通常并没有关于总体标准差的一个好的估计。在这种情形下，我们必须利用同一样本估计 μ 和 σ 两个未知参数。这属于 **σ 未知**（σ unknown）的情形。当利用 s 估计 σ 时，边际误差和总体均值的区间估计都是以 **t 分布**（t distribution）的概率分布为依据进行的。虽然 t 分布的数学推导是以假设抽样总体服从正态分布为依据，但是研究表明在许多总体分布显著偏离正态分布的情形下，利用 t 分布的效果还是相当不错的。当总体的分布不是正态分布时，在本节稍后我们给出了应用 t 分布的建议。

t 分布[㊀]是由一类相似的概率分布组成的分布族，某个特定的 t 分布依赖于称为**自由度**（degrees of freedom）

㊀ William Sealy Gosset 是牛津大学的一名数学研究生，笔名"student"，他在爱尔兰的都柏林为 Guinness Brewery 工作时在小容量材料和温度的试验中发现了 t 分布。

的参数。当自由度分别为1，2，3，…，时，有且仅有唯一的t分布与之相对应。随着自由度的增大，t分布与标准正态分布之间的差别变得越来越小。图8-4给出了自由度分别为10和20时的t分布与标准正态概率分布的关系。我们注意，随着自由度的增大，t分布的变异幅度减小，与标准正态分布也越来越相似。还注意到，t分布的均值为0。

我们给t加上下标以表明其在t分布上侧的面积。例如，就像我们用$z_{0.025}$表示在标准正态概率分布该值上侧的面积为0.025一样，我们用$t_{0.025}$表示在t分布该值上侧的面积为0.025。一般地，用记号$t_{\alpha/2}$表示在t分布中$t_{\alpha/2}$上侧的面积为$\alpha/2$（见图8-5）。

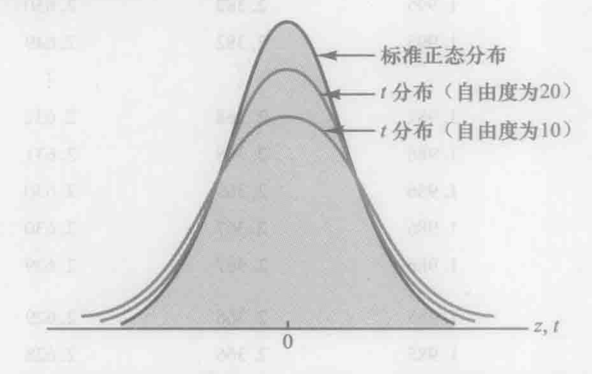


图8-4 比较自由度为10和20的t分布与标准正态分布

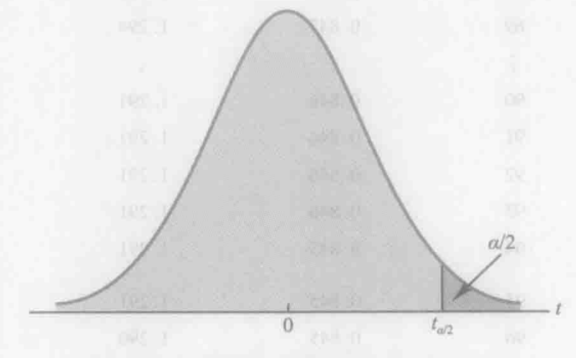


图8-5 上侧面积或概率为$\alpha/2$的t分布

表8-2 t分布表的部分值

自由度	上侧面积					
	0.20	0.10	0.05	0.025	0.01	0.005
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	0.978	1.638	2.353	3.182	4.541	5.841
4	0.941	1.533	2.132	2.776	3.747	4.604
5	0.920	1.476	2.015	2.571	3.365	4.032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1.397	1.860	2.306	2.896	3.355
9	0.883	1.383	1.833	2.262	2.821	3.250
⋮	⋮	⋮	⋮	⋮	⋮	⋮
60	0.848	1.296	1.671	2.000	2.390	2.660
61	0.848	1.296	1.670	2.000	2.389	2.659
62	0.847	1.295	1.670	1.999	2.388	2.657
63	0.847	1.295	1.669	1.998	2.387	2.656

（续）

自由度	上侧面积					
	0.20	0.10	0.05	0.025	0.01	0.005
64	0.847	1.295	1.669	1.998	2.386	2.655
65	0.847	1.295	1.669	1.997	2.385	2.654
66	0.847	1.295	1.668	1.997	2.384	2.652
67	0.847	1.294	1.668	1.996	2.383	2.651
68	0.847	1.294	1.668	1.995	2.382	2.650
69	0.847	1.294	1.667	1.995	2.382	2.649
⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	0.846	1.291	1.662	1.987	2.368	2.632
91	0.846	1.291	1.662	1.986	2.368	2.631
92	0.846	1.291	1.662	1.986	2.368	2.630
93	0.846	1.291	1.661	1.986	2.367	2.630
94	0.845	1.291	1.661	1.986	2.367	2.629
95	0.845	1.291	1.661	1.985	2.366	2.629
96	0.845	1.290	1.661	1.985	2.366	2.628
97	0.845	1.290	1.661	1.985	2.365	2.627
98	0.845	1.290	1.661	1.984	2.365	2.627
99	0.845	1.290	1.660	1.984	2.364	2.626
100	0.845	1.290	1.660	1.984	2.364	2.626
∞	0.842	1.282	1.645	1.960	2.326	2.576

注：附录B中的表B-2给出了更详细的t分布表。

附录B中表B-2是一张t分布表。表8-2是该表的一部分。表中的每一行分别与某一自由度的t分布相对应。例如，对自由度为9的t分布，$t_{0.025}=2.262$。类似地，对自由度为60的t分布，$t_{0.025}=2.000$。随着自由度的持续增大，$t_{0.025}$趋向于$z_{0.025}=1.96$ ⊖。实际上，标准正态分布的z值可以在t分布表中自由度为无穷的那一行（记做∞）找到。当自由度超过100时，自由度为无穷的那一行可以用于近似实际的t值。换句话讲，当自由度大于100时，标准正态的z值是对t值的一个好的近似。

8.2.1 边际误差和区间估计

在第8.1节，我们在σ已知的情形下给出总体均值的区间估计为

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

在σ未知的情形下，为了计算μ的区间估计，用样本标准差s估计σ，用t分布中的值$t_{\alpha/2}$代替$z_{\alpha/2}$。于是，边际误差为$t_{\alpha/2}s/\sqrt{n}$。利用该边际误差，当σ未知时，总体均值区间估计的一般公式如下所示。

总体均值的区间估计：σ未知情形

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{8-2}$$

式中，s为样本标准差；$1-\alpha$为置信系数，自由度为$n-1$的t分布中，$t_{\alpha/2}$上侧的面积恰好等于$\alpha/2$。

由于用s作为总体标准差σ的估计值，所以在式（8-2）中与t值对应的自由度为$n-1$。样本标准差的公式为

⊖ 随着自由度的增大，t分布越来越接近于正态分布。

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

自由度是计算 $\sum (x_i - \bar{x})^2$ 时所用到的信息中独立信息的个数。在计算 $\sum (x_i - \bar{x})^2$ 时，用到如下 n 条信息：$(x_1 - \bar{x})$，$(x_2 - \bar{x})$，…，$(x_n - \bar{x})$。在第3.2节中，我们已经证明了对于任何数据集 $\sum (x_i - \bar{x}) = 0$。因此，$(x_i - \bar{x})$ 中只有 $n-1$ 项是独立的，即如果我们知道了 $n-1$ 个值，则由所有 $(x_i - \bar{x})$ 值之和为0，可以确定余下的值。于是，与 $\sum (x_i - \bar{x})^2$ 所联系的自由度的个数为 $n-1$，因此式（8-2）中 t 分布的自由度为 $n-1$。

为演示 σ 未知的情形下区间估计的方法，我们设计一项研究用于估计美国家庭信用卡债务的总体均值。由 $n=70$ 个家庭组成样本，他们的信用卡余额数据如表8-3所示。在这种情况下，没有关于总体标准差 σ 的先验估计。因此，必须利用样本数据同时估计总体均值和总体标准差。利用表8-3中数据，计算得到样本均值 $\bar{x} = 9\,312$ 美元和样本标准差 $s = 4\,007$ 美元。在95%的置信水平和自由度为 $n-1=69$ 的情形下，查表8-2可得 $t_{0.025}$ 的值。位于自由度为69的行和上侧概率为0.025的列上的 t 值，$t_{0.025} = 1.995$。

表8-3 样本中70个家庭的信用卡余额数据

9 430	14 661	7 159	9 071	9 691	11 032
7 535	12 195	8 137	3 603	11 448	6 525
4 078	10 544	9 467	16 804	8 279	5 239
5 604	13 659	12 595	13 479	5 649	6 195
5 179	7 061	7 917	14 044	11 298	12 584
4 416	6 245	11 346	6 817	4 353	15 415
10 676	13 021	12 806	6 845	3 467	15 917
1 627	9 719	4 972	10 493	6 191	12 591
10 112	2 200	11 356	615	12 851	9 743
6 567	10 746	7 117	13 627	5 337	10 324
13 627	12 744	9 465	12 557	8 372	
18 719	5 742	19 263	6 232	7 445	

我们利用式（8-2）计算信用卡余额的总体均值的区间估计

$$9\,312 \pm 1.995\left(\frac{4\,007}{\sqrt{70}}\right) \qquad 9\,312 \pm 955$$

总体均值的点估计为9 312美元，边际误差为955美元，95%的置信区间为9 312 − 955 = 8 357美元至9 312 + 955 = 10 267美元。于是，我们有95%的把握认为：对于全部家庭的总体，信用卡余额的均值介于8 357 ~ 10 267美元。

在附录8A、附录8B和附录8C中，描述了如何利用Minitab、Excel和StatTools求总体均值的置信区间。在家庭信用卡余额的研究中，Minitab区间估计程序的输出结果见图8-6。70个家庭中信用卡余额的样本均值为9 312美元，样本标准差为4 007美元，均值的标准误差的估计为479美元，95%置信区间为（8 357，10 267）。

```
Variable      N     Mean    StDev   SE Mean          95% CI
NewBalance   70     9312     4007       479    (8357, 10267)
```

图 8-6

8.2.2 应用中的建议

如果总体服从正态分布，式（8-2）所给出的置信区间是精确的，并且适用于任何样本容量。如果总体不服从正态分布，则式（8-2）所给出的置信区间是近似的。在这种情形下，近似的程度依赖于总体的分布和样本容量。

在绝大部分应用中，当利用式（8-2）建立总体均值的区间估计时，样本容量 $n \geqslant 30$ 已经足够大。然而，如果总体的分布是严重偏斜或者包含异常点时，绝大部分统计学家建议将样本容量增加到50或者更大。[⊖]如果总体

⊖ 如果总体的分布是严重偏斜或者包含异常点时，需要更大的样本容量。

的分布不是正态分布但是大致对称，则在样本容量为15时便能得到置信区间一个好的近似。仅当分析人员坚信或者愿意假设总体的分布至少是近似正态时，才可以在更小的样本容量下使用式（8-2）。

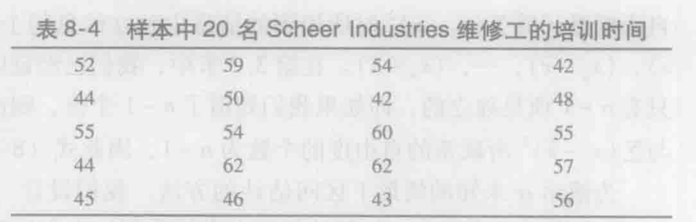
8.2.3 利用小样本

在下面的例子中，我们在小样本容量下考虑如何构建总体均值的区间估计。正如所强调的那样，在断定区间估计方法能否给出可以接受的结果时，对总体分布的了解至关重要。

Scheer Industries考虑使用一种新型的计算机辅助程序来培训员工维修机器。为了对这种程序有一个全面的评估，生产负责人要求对维修工完成该计算机辅助培训所需时间的总体均值进行估计。

选取20名员工组成一个样本，样本中的每名员工都完成了培训计划。每名维修工培训所用的时间数据如表8-4所示（单位：天）。图8-7是样本数据的直方图。根据直方图，我们认为总体的分布是怎样的呢？首先，样本数据不能支持我们得出总体服从正态分布的结论。但是，我们也没有发现任何偏斜或者异常点方面的证据。因此，根据上一节的提示，我们的结论是：对于这20名员工组成的样本，以t分布为依据进行区间估计是可行的。

表8-4 样本中20名Scheer Industries维修工的培训时间



52	59	54	42
44	50	42	48
55	54	60	55
44	62	62	57
45	46	43	56

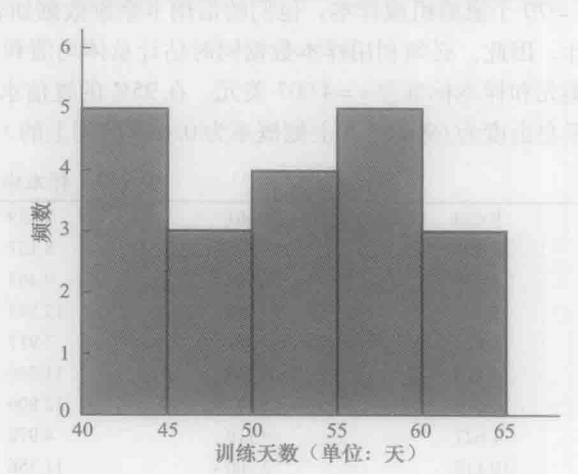


图8-7 在Scheer Industries样本中培训时间的直方图

计算样本均值和样本标准差，得到：

$$\bar{x}=\frac{\sum x_i}{n}=\frac{1\,030}{20}=51.5(\text{天}) \qquad s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}=\sqrt{\frac{889}{20-1}}=6.84(\text{天})$$

对于95%的置信水平和自由度$n-1=19$，查附录B中表B-2，得到$t_{0.025}=2.093$。式（8-2）给出总体均值的区间估计为

$$51.5\pm2.093\left(\frac{6.84}{\sqrt{20}}\right) \qquad 51.5\pm3.2$$

总体均值的点估计为51.5天。边际误差为3.2天，95%的置信区间为$51.5-3.2=48.3$到$51.5+3.2=54.7$。

利用样本数据的直方图了解总体的分布时，并不总能得到令人信服的结论，但是在许多情形下，它给出了仅有的可利用的信息。通常，利用直方图结合分析人员的部分判断，来确定是否可以利用式（8-2）进行区间估计。

8.2.4 区间估计方法小结

我们给出了总体均值的区间估计的两种方法：对于σ已知的情形，式（8-1）利用标准正态分布和σ来计算边际误差并进行区间估计；对于σ未知的情形，式（8-2）利用样本标准差s和t分布来计算边际误差并进行区间估计。

图8-8是对这两种情形下区间估计方法的小结。在绝大部分应用中，样本容量$n\geqslant30$已经足够大。然而，如果总体服从或者近似服从正态分布，则可以采用更小的样本容量。对于σ未知情形，如果总体的分布是严重偏斜或者包含异常点时，建议将样本容量增加到$n\geqslant50$。

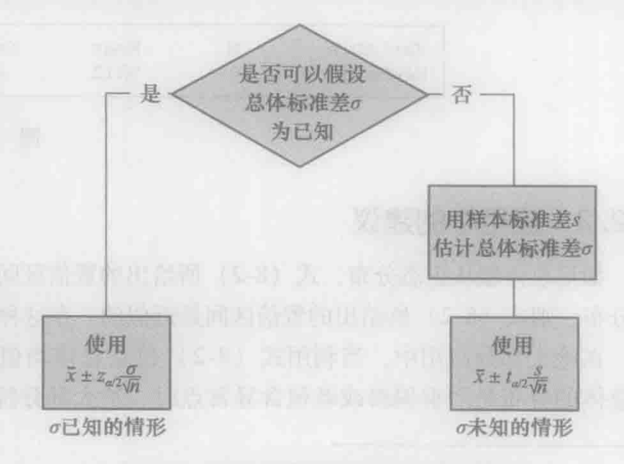


图8-8 总体均值区间估计的方法小结

注释和评论

1. 当σ已知时，对所有容量为n的样本边际误差为$z_{\alpha/2}(\sigma/\sqrt{n})$是固定的，并且是相同的。当$\sigma$未知时，边际误差$t_{\alpha/2}(s/\sqrt{n})$随样本的变化而变化。这是由于样本标准差$s$依赖于抽取的样本。$s$越大边际误差越大，$s$越小边际误差越小。
2. 如果总体分布是偏斜的，置信区间的估计将如何变化？当总体的分布是偏向右侧时，数据中大的值将分布拉向右侧。在这种偏斜下，样本均值$\bar{x}$和样本标准差s是正相关的。较大的s与较大的$\bar{x}$值相对应。于是，当$\bar{x}$大于总体均值时，s将大于σ。这种偏斜使得边际误差$t_{\alpha/2}(s/\sqrt{n})$大于σ已知情形的边际误差为$z_{\alpha/2}(\sigma/\sqrt{n})$。与使用$\sigma$的真值相比，大的边际误差所对应的置信区间将以更大的可能性将总体均值μ包括在内。但是，当样本均值$\bar{x}$小于总体均值时，$\bar{x}$和s的相关性使得边际误差较小。在这种情形下，与σ已知从而可用相比，较小的边际误差所对应的置信区间将以更小的可能性将总体均值μ包括在内。出于这种原因，我们建议在总体分布严重偏斜时采用更大的样本容量。

练习

方法

12. 下面各种情形下中，t值为多少？
   a. 自由度为12，上侧面积为0.025。
   b. 自由度为50，下侧面积为0.05。
   c. 自由度为30，上侧面积为0.01。
   d. 自由度为25，两个t值之间面积为90%。
   e. 自由度为45，两个t值之间面积为95%。

14. 一个简单随机样本由$n=54$项组成，样本均值为22.5，样本标准差为4.4。
   a. 求总体均值的90%置信区间。
   b. 求总体均值的95%置信区间。
   c. 求总体均值的99%置信区间。
   d. 当置信水平增大时，边际误差和置信区间如何变化？

应用

16. 40家企业的债券组成一个样本，到期年数和获利的数据存放在名为CorporateBonds的文件中（《巴伦周刊》，2012年4月2日）。
   a. 求企业债券到期年数的样本均值和样本标准差。
   b. 求到期年数的总体均值的95%置信区间估计。
   c. 求企业债券获利的样本均值和样本标准差。
   d. 求企业债券获利的总体均值的95%置信区间估计。

☆18. 年龄越大的人通常越难找工作。美国退休协会（AARP）报道了年龄55岁及以上的人找工作时所耗费的时间（单位：周）。数据存放在名为JobSearch的文件中（*AARP Bulletin*，2008年4月）。
   a. 求年龄55岁及以上的人找工作时所耗费时间的总体均值的点估计。
   b. 在95%的置信水平下，边际误差是多少？
   c. 估计均值的95%置信区间。
   d. 讨论样本数据的偏斜程度。若再次展开研究，你有何建议？

20. 你所喜爱的电视节目是否经常被广告打断？美国全国广播公司财经频道（CNBC）统计了半个小时的情景喜剧时间段中，节目播出的平均时间长度（CNBC，2006年2月23日）。调查数据如下（单位：分钟）：

21.06	22.24	20.62	21.66	21.23
23.86	23.82	20.30	21.52	21.52
21.91	23.14	20.02	22.20	21.20
22.37	22.19	22.34	23.36	23.44

假设总体近似服从正态分布，试求在半个小时的情景喜剧时间段中，节目播出时间平均长度的点估计及95%的置信区间估计。

22. 2009年4月，迪士尼的电影《蒙塔娜·汉娜》恰逢复活节的周末上映。在周末的三天当中，成为最具票房吸引力的电影（《华尔街日报》，2009年4月13日）。由25家剧院组成一个样本，它们的票房收入如下（单位：美元）

20 200	10 150	13 000	11 320	9 700
8 350	7 300	14 000	9 940	11 200
10 750	6 240	12 700	7 430	13 500
13 900	4 200	6 750	6 700	9 330
13 185	9 200	21 400	11 380	10 800

   a. 求每家剧院票房收入均值的95%置信区间。解

释这个结果。

b. 电影票的价格是每张7.16美元，试估计每家影院平均观影人数。

c. 该部影片已经在3 118家剧院上映。估计观看电影《蒙塔娜·汉娜》的观众总人数，并估计周末三天当中的票房总收入。

8.3 样本容量的确定

在前文的应用建议中，当总体不服从正态分布时，我们重点讨论了样本容量在得到好的置信区间中的作用。在本节，我们重点介绍与样本容量有关的另一个问题，说明如何确定足够的样本容量以达到所希望的边际误差[1]。为理解这一方法如何实施的，我们回到第8.1节σ已知的情形。根据式（8-1），区间估计为

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$z_{\alpha/2}(\sigma/\sqrt{n})$是边际误差。于是，我们看到：值$z_{\alpha/2}$、总体标准差$\sigma$和样本容量$n$共同确定了边际误差。一旦选择了置信系数$1-\alpha$，$z_{\alpha/2}$就确定了。然后，如果$\sigma$的值已知，我们就可以确定达到希望的边际误差所需的样本容量n。以下是用于计算所需样本容量n的公式。

令E代表希望达到的边际误差

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

解出$\sqrt{n}$，得到

$$\sqrt{n} = \frac{z_{\alpha/2}\cdot\sigma}{E}$$

将上式两边平方，得到样本容量的如下表达式。

总体均值区间估计中的样本容量[2]

$$n = \frac{(z_{\alpha/2})^2\cdot\sigma^2}{E^2} \tag{8-3}$$

在给定的置信水平下，这一样本容量能达到希望的边际误差。

在式（8-3）中，E值是使用者可接受的边际误差，$z_{\alpha/2}$可直接由区间估计中所用到的置信水平确定。虽然不同的使用者有各自的偏好，但95%的置信水平仍是最常见的选择（$z_{0.025}=1.96$）。

最后，式（8-3）要求总体标准差σ的值必须是已知的。但是，即使σ是未知的，只要我们能给出σ的初始值或计划值，仍可以使用式（8-3）。在实践中，可选择以下方法之一来确定σ的值[3]。

1. 根据以前研究中的数据计算总体标准差的估计值作为σ的计划值。
2. 利用实验性研究，选取一个初始样本，以初始样本的标准差作为σ的计划值。
3. 对σ值进行判断或最优猜测。例如，我们可以分别估计总体的最大值和最小值，两者之差是对数据极差的估计。一般建议将极差除以4做为标准差σ的粗略估计，从而最终得到一个可以接受的σ的计划值。

下面，我们举例说明如何利用式（8-3）来确定样本容量。对美国汽车租赁成本的已有调查研究发现，租赁一辆中型汽车的平均费用大约为每天55美元。假定该项研究的组织者想要进行一项新的调查，对在美国一辆中型汽车的租赁费用的总体均值进行估计。在新研究的设计中，当项目负责人估计每天租赁费用的总体均值时，设定置信水平为95%，边际误差为2美元。

该负责人已设定边际误差$E=2$，在95%的置信水平下$z_{0.025}=1.96$，因此在计算所需的样本容量时，仅需要已知总体标准差σ的计划值。此时，分析人员回顾先前研究中的样本数据发现，每天租赁费用的样本标准差为9.65

---

[1] 如果在抽样之前就选定了所希望的边际误差E，则本节的方法能用于确定的必要的样本容量，使得所要求的边际误差得到满足。
[2] 式（8-3）所给出的是满足所要求的边际误差所需的最优样本容量。但是，可以根据分析者的判断来确定是否将最终增大样本容量。
[3] 在确定样本容量之前必须设定总体标准差σ的计划值。这里讨论了得到σ的计划值的三种方法。

美元。将9.65作为σ的计划值，得到

$$n = \frac{(z_{\alpha/2})^2 \cdot \sigma^2}{E^2} = \frac{(1.96)^2 \times (9.65)^2}{2^2} = 89.43$$

因此，在新研究中为达到项目负责人所指定的边际误差2美元，样本中至少应该选取89.43笔中型汽车租赁业务。当计算得到的样本容量不是整数时，建议使用取下一位整数为样本容量[⊖]。因此，建议样本容量为90笔中型汽车租赁业务。

练习

方法

☆24. 估计一组数据的极差为36，试求：

a. 总体标准差的计划值是多少？

b. 在95%的置信水平下，当边际误差为3时应选取多大的样本容量？

c. 在95%的置信水平下，当边际误差为2时应选取多大的样本容量？

应用

26. 据美国能源信息协会（US EIA）报道，普通汽油的平均价格为每加仑3.94美元（US EIA网站，2006年2月23日）。US EIA每周更新其对汽油均价的估计。假定每加仑普通汽油价格的标准差为0.25美元，若在95%置信度下，按如下边际误差报道油价时，应该选取多大的样本容量？

a. 所期望的边际误差为0.10美元。

b. 所期望的边际误差为0.07美元。

c. 所期望的边际误差为0.05美元。

28. ShareBuilder是一名退休计划建议师，根据他和哈里斯互动进行的一项在线调查报告，女企业主当中有60%的人不能确保自己目前为退休所做的储蓄是足够的（SmallBiz，2006年冬）。假定我们想要做一项后续研究，确定女企业主每年为退休所做的储蓄额有多少，并且希望总体均值区间估计的边际误差为100美元。取标准差的计划值为1 100美元，则在下列情形下，建议选取多大的样本容量？

a. 构建储蓄额均值的90%的置信区间。

b. 构建储蓄额均值的95%的置信区间。

c. 构建储蓄额均值的99%的置信区间。

d. 当边际误差固定不变时，增大置信水平将引起样本容量如何变化？你建议采用99%的置信区间吗？为什么？

30. 在过去几年中，人们特别是年轻人表现出不愿驾驶的趋势。2001~2009年，16~34岁年龄段人中每人每年旅行的车英里数从10 300下降到7 900（U. S. PIRG和教育基金网站，2012年4月6日）。假设2009年的标准差为2 000英里。假定你要进行一个调查，对目前16~34岁年龄段人中每人每年的车－英里进行区间估计，若边际误差为100英里，则目前的调查中应采用多大的样本容量？取95%的置信水平。

8.4 总体比率

在本章前言中，我们给出总体比率p的区间估计的一般形式为

$$\bar{p} \pm \text{边际误差}$$

在计算区间估计的边际误差时，$\bar{p}$的抽样分布至关重要。

在第7章，我们已经证明了当$np \geq 5$和$n(1-p) \geq 5$时$\bar{p}$的抽样分布近似服从正态分布，图8-9给出了$\bar{p}$的抽样分布的正态近似。$\bar{p}$的抽样分布的均值是总体比率p，$\bar{p}$的标准差为

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{8-4}$$

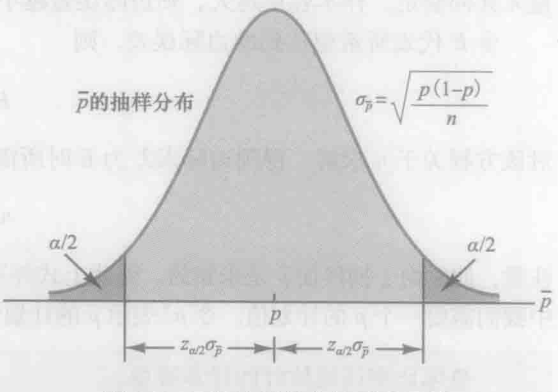


图8-9 $\bar{p}$的抽样分布的正态近似

⊖ 式（8-3）所给出的是满足所希望的边际误差所需要的最小样本容量。如果计算得到的样本容量不是整数，则建议使用取下一位整数为样本容量，这时的边际误差比所要求的边际误差略小。

由于$\bar{p}$的抽样分布服从正态分布，如果在总体比率的区间估计中选择$z_{\alpha/2}\sigma_{\bar{p}}$作为边际误差，则所得到的区间中有100（1－α)%将真实的总体比率包含在内。但是，由于p未知，在计算边际误差时不能直接使用$\sigma_{\bar{p}}$。于是，我们需要估计p，用$\bar{p}$代替p。在总体比率的区间估计中，边际误差为

$$\text{边际误差} = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8-5}$$

总体比率的区间估计的一般公式如下。

总体比率的区间估计[⊖]

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8-6}$$

式中，$1-\alpha$为置信系数；$z_{\alpha/2}$为标准正态分布上侧面积为$\alpha/2$时的z值。

下面的例子说明了如何计算总体比率的边际误差和区间估计。美国对900名女子高尔夫运动员进行了一项全国性调查，以便掌握女子高尔夫运动员如何看待她们在高尔夫球场所受到的待遇。调查显示，有396名女子高尔夫运动员对开球时间的合理性感到满意。于是，女子高尔夫运动员中对开球时间的合理性感到满意的总体比率的点估计为396/900＝0.44。根据式（8-6）并取置信水平为95%，可得

$$\bar{p} \pm z_{0.025}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$0.44 \pm 1.96\sqrt{\frac{0.44(1-0.44)}{900}}$$

$$0.44 \pm 0.0324$$

因此，边际误差为0.0324，总体比率的95%置信水平的区间估计为（0.4076，0.4724)。若用百分数来表示，调查结果使我们可以认为在95%的置信水平下全部女子高尔夫运动员中对开球时间的合理性感到满意的比率在40.76%～47.24%。

8.4.1 样本容量的确定

下面我们进一步考虑，在给定的精度水平下，应选取多大的样本容量来估计总体比率。总体比率的区间估计中确定样本容量的原理与第8.3节中总体均值的区间估计中样本容量的确定完全类似。

在本节开始，与总体比率的区间估计相联系的边际误差为$z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}$，它由$z_{\alpha/2}$、样本比率$\bar{p}$和样本容量$n$共同确定。样本容量越大，则边际误差越小，精度越高。

令E代表所希望达到的边际误差，则

$$E = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

对该方程关于n求解，得到边际误差为E时所需样本容量的公式。

$$n = \frac{(z_{\alpha/2})^2\bar{p}(1-\bar{p})}{E^2}$$

注意，此时由于抽样前$\bar{p}$是未知的，因此上式并不能用于计算达到预期的边际误差所需要的样本容量。于是，计算中我们需要一个$\bar{p}$的计划值。令p^*表示$\bar{p}$的计划值，则式（8-7）可用于计算边际误差为E时所需的样本容量。

总体比率区间估计的样本容量

$$n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2} \tag{8-7}$$

⊖ 在求比率的区间估计时，边际误差是$z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}$。

实践中，可选择如下方法来确定计划值p^*。

(1) 用以前相同或类似样本的样本比率来代替。

(2) 利用实验性的研究，选取一个初始样本，以该样本的样本比率作为计划值p^*。

(3) 使用判断或最优猜测作为p^*值。

(4) 如果上述方法均不适用，则取计划值$p^*=0.5$。

我们仍回到对女子高尔夫运动员的调查中去。假设公司想通过一项新的调查来估计目前女子高尔夫运动员中对开球时间的合理性感到满意的总体比率。在95%的置信水平和边际误差为0.025的条件下，调查组织者对总体比率进行估计时应选取多大的样本容量？由$E=0.025$和$z_{\alpha/2}=1.96$，我们还需要计划值p^*方能回答所需样本容量的问题。将上一次的调查结果$\bar{p}=0.44$作为计划值p^*，根据式(8-7)，有

$$n=\frac{(z_{\alpha/2})^2p^*(1-p^*)}{E^2}=\frac{(1.96)^2\times0.44\times(1-0.44)}{0.025^2}=1\,514.5$$

因此，为达到所要求的边际误差，样本容量至少应为1 514.5人。我们取下一个整数，则为达到所要求的边际误差，建议由1 515名女子高尔夫运动员组成一个样本。

在选择计划值p^*时，我们所推荐使用的第四种方法为取$p^*=0.5$。在没有其他信息可用时，p^*通常取此值。式(8-7)的分子表明样本容量与数值$p^*(1-p^*)$成正比，$p^*(1-p^*)$的值越大则样本容量越大。表8-5给出了$p^*(1-p^*)$的一些可能值，当$p^*=0.5$时，$p^*(1-p^*)$的值达到最大。因此，当我们不能确定计划值取何值时，由$p^*=0.5$所计算得到的样本容量最大。实际上，出于保险的原则，我们建议使用最大的样本容量。如果能够证明样本比率不是计划值0.5，则边际误差比预期的要小。因此，取$p^*=0.5$计算出的样本容量足以保证达到希望的边际误差。

表8-5 $p^*(1-p^*)$的一些可能值

p^*	$p^*(1-p^*)$	
0.1	0.1×0.9=0.09	
0.3	0.3×0.7=0.21	
0.4	0.4×0.6=0.24	
0.5	0.5×0.5=0.25	← $p^*(1-p^*)$的最大值
0.6	0.6×0.4=0.24	
0.7	0.7×0.3=0.21	
0.9	0.9×0.1=0.09	

在对女子高尔夫运动员调查的例子中，当取计划值$p^*=0.5$时样本容量为

$$n=\frac{(z_{\alpha/2})^2p^*(1-p^*)}{E^2}=\frac{(1.96)^2\times0.5\times(1-0.5)}{0.025^2}=1\,536.6$$

因此建议采用稍大的样本容量——选取1 537名女子高尔夫运动员组成样本。

注释和评论

在估计总体比率时，边际误差几乎总是小于或等于0.10。在诸如盖洛普和哈里斯所进行的全美范围的公众民意调查中，边际误差一般为0.03或0.04。根据这样的边际误差，由式(8-7)所给出的样本容量一般都是足够大的，满足条件$np\geqslant5$和$n(1-p)\geqslant5$，从而可以用正态分布近似$\bar{x}$的抽样分布。

练习

方法

32. 800个个体组成一个简单随机样本，样本比率为$\bar{p}=0.7$。

a. 求总体比率的90%置信区间。

b. 求总体比率的95%置信区间。

34. 在95%的置信水平下，边际误差为0.03时，为了估计总体比率，应采用多大的样本容量？假设不能根据历史数据得到计划值p^*。

应用

36. 据CNBC的统计报告，没上保险的机动车的数量令人吃惊（CNBC，2006年2月23日）。与CNBC的报道一致，抽样结果显示在200辆机动车中有46辆没有上保险。

a. 试求机动车没有上保险的比率的点估计。

b. 建立总体比率的95%的置信区间。

38. 据汤姆森金融公司称，到2006年1月25日为止，

大部分公司公布的利润都比估计的要高（《商业周刊》，2006年2月6日）。由162家公司组成的一个样本显示，有104家的利润比估计的高，29家的利润与估计相符，29家的利润低于估计。

a. 求利润低于估计的公司所占比率的点估计。

b. 试求边际误差，并给出利润高于估计的公司所占比率的95%的置信区间。

c. 要达到预期的边际误差0.05，需要多大的样本容量？

40. 企业多年来一直与不断增长的医疗保健费用作斗争。最近，由于较低的通货膨胀和雇员支付较大比例的医疗保健福利金，使得医疗保健费用的增长速度放缓。最近，Mercer的一项调查显示，2009年美国52%的雇主要求员工分担更多的医疗保健费用（《商业周刊》，2009年2月16日）。假设，调查中样本是由800家公司组成，计算边际误差，并在95%的置信水平下，建立2009年雇主要求员工分担更多的医疗保健费用的公司所占比率的置信区间。

42. 6月，对491名潜在选民就总统大选进行了一项民意调查。本次民意调查的主要目的是，估计潜在选民中对每位候选人的支持率。假设计划值 $p^* = 0.50$，取置信水平为95%。

a. 对 $p* = 0.50$，6月的民意调查中计划的边际误差为多少？

b. 临近11月大选前夕，希望达到更高的精确度和更小的边际误差。假设对总统大选期间所进行的调查要求满足如下边际误差，则每次调查中建议选取多大的样本容量？

调查时间	边际误差	调查时间	边际误差
9月	0.04	11月初	0.02
10月	0.03	大选前一天	0.01

小　结

本章我们给出了对总体比率和总体均值进行区间估计的方法。点估计量可能是也可能不是总体参数的好的估计。利用区间估计可以对估计的精确程度予以度量。总体均值和总体比率的区间估计都采用如下形式：点估计±边际误差。

在两种情形下，我们分别介绍了总体均值的区间估计。在 σ 已知情形下，抽样前利用历史数据以及其他信息对 σ 进行了估计。于是，在根据新的样本数据进行分析时，可以假设 σ 已知。在 σ 未知情形下，利用样本数据同时估计总体均值和总体标准差。最终究竟选择哪一种区间估计方法，依赖于分析人员认为何种方法能给出对 σ 的最佳估计。

在 σ 已知情形下，基于假设的 σ 值并利用标准正态分布进行区间估计。在 σ 未知情形下，基于样本标准差 s 并利用 t 分布进行区间估计。在这两种情形下，所得到的区间估计的质量取决于总体的分布和样本容量。如果总体服从正态分布，则即使在小样本时这两种情形所得到的区间估计都是精确的。如果总体不服从正态分布，则所得到的区间估计都是近似的。样本容量越大，近似程度越好。与正态分布相比，总体的偏斜程度越大，要想得到更好的近似，需要的样本容量也越大。第8.1节和第8.2节中，对达到好的近似应使用多大的样本容量提出了实践应用的建议。在大多数情况下，30或者更大的样本容量将保证能给出置信区间一个好的近似。

总体比率的区间估计的一般形式为：$\bar{p}$ ±边际误差。实际应用中，对总体比率进行区间估计时，通常要求样本容量为大样本。因此，区间估计是依据标准正态分布进行的。

在开展一项抽样计划之前，通常先设定一个希望达到的边际误差。我们说明如何选择足够大的样本容量来达到所希望的精度。

关键术语

interval estimate 区间估计 总体参数估计值的一个区间，确信该区间将参数值纳入其中。在本章中，区间估计的形式是：点估计±边际误差。

margin of error 边际误差 为了建立总体参数的区间估计，从点估计值加上或减去的值。

σ known σ 已知 在进行抽样之前，若历史数据或者其他信息给出了总体标准差的一个好的近似，称这种情形为 σ 已知。在区间估计中利用这个已知的 σ 值计算边际误差。

confidence level 置信水平 与区间估计相联系的置信度。例如，在使用区间估计方法得到的全部区间中，如果有95%的区间包含总体参数在内，则称区

间估计是在95%的置信水平下构建的。

confidence coefficient **置信系数** 用小数形式表示的置信水平。比如，95%置信水平的置信系数为0.95。

confidence interval **置信区间** 区间估计的另一种叫法。

σ unknown **σ未知** 在进行抽样之前，更普遍的情形是没有信息可用于估计总体标准差σ。我们称这种情形为σ未知。这时，在区间估计中计算边际误差时用的是样本标准差s。

t distribution **t分布** 一类概率分布。当总体标准差σ未知而用样本标准差s对其进行估计时，该分布用于建立总体均值的区间估计。

degrees of freedom **自由度** t分布的参数，当t分布用于计算总体均值的区间估计时，其自由度为n−1，其中n是样本容量。

重要公式

总体均值的区间估计：σ已知

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad (8\text{-}1)$$

总体均值的区间估计：σ未知

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \quad (8\text{-}2)$$

总体均值的区间估计中所需的样本容量

$$n = \frac{(z_{\alpha/2})^2 \cdot \sigma^2}{E^2} \quad (8\text{-}3)$$

总体比率的区间估计

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad (8\text{-}6)$$

总体比率的区间估计中所需的样本容量

$$n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2} \quad (8\text{-}7)$$

补充练习

44. 对54名贴现票据经纪人的一个抽样调查显示，交易股价50美元的股票手续费的均值为33.77美元/百股（*AAII Journal*，2006年2月）。这样的调查每年进行，根据可得的历史数据，假定已知总体标准差为15美元。
   a. 利用样本数据，求与95%置信区间相对应的边际误差。
   b. 当交易股价50美元的股票时，求手续费的均值的95%置信区间。

46. 美国50岁及以上年龄的人口达9 200万，他们掌握了整个可支配收入的50%（*AARP Bulletin*，2008年3月）。据美国退休人员协会（AARP）估计，在这一年龄段食堂和外卖的年人均消费为1 873美元。假定样本由80名民众组成，并且样本标准差为550美元。
   a. 当置信度为95%时，边际误差为多大？
   b. 求用于食堂和外卖上的消费额的总体均值的95%置信区间。
   c. 估计美国50岁及以上年龄的人用于食堂和外卖上的消费是多少？
   d. 如果食堂和外卖上的消费额是右偏的，那么你预计消费额的中位数是大于还是小于1 873美元？

☆48. 美国合众国航空的大量研究表明，鼓励那些因频繁飞行而享有里程积分的乘客通过在线的方式确定里程和安排飞行奖励，这可以从根本上节省开支（*US Airways Attache*，2003年2月）。一项研究搜集了通过电话确定里程和安排飞行奖励时所用时间的数据，通过电话安排150次飞行奖励所需时间的数据（单位：分）组成一个样本，存放在名为Flights的数据集中。请借助Minitab或者Excel回答下列问题。
   a. 求通过电话方式安排飞行奖励所需时间的样本均值。
   b. 求通过电话方式安排飞行奖励所需时间的总体均值的95%置信区间。
   c. 假设一家电话订票代理处每天工作7.5个小时，则预计一天中这家电话订票代理处能够处理多少笔飞行奖励？
   d. 讨论：为什么说这一信息支持美国合众国航空采用在线系统以降低成本的计划？

50. 对某种型号汽车进行了一项里程测试。如果希望达到的边际误差为1公里/加仑，则测试中应选取多少辆汽车？取置信水平为98%，并假设里程预测试表明标准差为2.6英里/加仑。

52.《商业周刊》的年度薪酬调查（Annual Pay Survey）

公布了首席执行官（CEO）的年薪及奖金的数据。在一个初始样本（单位：千美元）中，标准差为675美元。如果我们想要对年薪和奖金的总体均值进行估计，设定边际误差为100 000美元，那么应由多少名CEO组成样本？（注意：由于数据单位是千美元，所以边际误差 $E=100$）。取置信水平为95%。

54. 《今日美国》/CNN/Gallup对369名在职父母进行的一项调查显示，他们当中有200人承认由于工作的原因而与子女相处时间太少。
    a. 对于由在职父母组成的总体，求由于工作的原因而与子女相处时间太少的父母所占比率的点估计。
    b. 当置信水平为95%时，边际误差为多少？
    c. 对于由在职父母组成的总体，求由于工作的原因而与子女相处时间太少的父母所占比率的95%置信区间。

56. 在2005年全美最为繁忙的机场当中，辛辛那提北肯塔基国际机场的准时到达率位居全美第二（*The Cincinnati Enquirer*，2006年2月3日）。假设由550个航班组成一个样本，其中有455个航班准时到达。
    a. 试求机场准时到达率的点估计。
    b. 构建2005年机场航班准时到达率的95%置信区间。

58. 一家著名的银行信用卡公司想要估计：在月末有余额并引起利息费用的信用卡持卡人所占比率。假定希望达到的边际误差为0.03，取置信水平为98%。
    a. 如果预期持卡人中大约有70%在月末有余额，则应该抽取多大容量的样本？
    b. 如果不能指定总体比率的计划值，则应该选取多大容量的样本？

60. 航空公司时刻表和价格是商务旅行者选择航班时要考虑的重要因素，但是《今日美国》的调查发现，商务旅行者将航空公司的经常性旅客优惠政策列为最重要的因素。在一个由 $n=1\,993$ 名商务旅行者组成的样本中，有618人将航空公司的经常性旅客优惠政策列为最重要的因素。
    a. 求商务旅行者总体中，在选择航班时将航空公司的经常性旅客优惠政策列为最重要的因素的人所占比率的点估计。
    b. 求总体比率的95%置信区间估计。
    c. 当置信水平为95%，边际误差为0.01时，应选取多大的样本容量？你认为《今日美国》的调查能达到这种精度吗？为什么？

案例8-1 YOUNG PROFESSIONAL杂志

Young Professional 杂志面向的读者是刚毕业的大学生，这些大学毕业生正处在他们职业生涯的头10年。在过去两年中，杂志的发行相当成功。目前出版商感兴趣的问题是如何扩大杂志的广告客户基础。潜在的广告客户不断地询问有关订阅 *Young Professional* 杂志的订阅人特征和订阅人兴趣方面的问题。为了收集这方面的信息，杂志进行了一次调查以便获取订阅人的概况资料。调查结果将用于帮助杂志选择感兴趣的文章并向做广告的人提供订阅人的概况。你身为杂志的一名新雇员，要求帮助分析调查结果。

如下是部分调查问题：

1. 您的年龄？
2. 您是：男性________　女性________。
3. 未来两年您是否有购买房产的计划？是________　否________。
4. 除去家庭投资之外，您或者您家庭的成员进行金融投资的总金额大约是多少？
5. 去年您共进行了多少次的股票/债券/互助基金交易？
6. 您家中有互联网宽带接入吗？是________　否________。
7. 您去年的家庭总收入是多少？
8. 您有子女吗？是________　否________。

这些问题的回答存放在光盘名为Professional的文件中。表8-6列示的是文件的一部分——包括调查中前5个受访者的回答。

表 8-6 *YOUNG PROFESSIONAL* 杂志的部分调查结果

年龄	性别	购买房产	投资额（美元）	交易次数	宽带接入	家庭收入（美元）	子女
38	女	否	12 200	4	是	75 200	有
30	男	否	12 400	4	是	70 300	有
41	女	否	26 800	5	是	48 200	无
28	女	是	19 600	6	否	95 300	无
31	女	是	15 100	5	否	73 300	有
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

管理报告

准备一份管理报告对调查结果进行汇总。除了统计汇总之外，讨论杂志应该如何利用这些结果来吸引需要登广告的人。你还可以建议刊物的编辑如何利用这些调查结果来判断读者对哪些话题更感兴趣。你的报告应该陈述如下问题，但又不必仅局限于这些领域。

1. 利用恰当的描述统计量对数据进行汇总。
2. 求订阅人年龄均值和订阅人家庭收入均值的置信区间，取置信度为95%。
3. 求订阅人家中有宽带接入的比率的95%置信区间。求有子女的订阅人所占比率的95%置信区间。
4. 对在线代理商而言，*Young Professional* 杂志可能是一个好的广告途径吗？根据统计数据判断你的结论。
5. 对销售幼儿教育软件和计算机游戏的企业来说，该杂志是刊登广告的好地方吗？
6. 谈谈你认为 *Young Professional* 杂志的读者会对哪些类型的文章感兴趣。

案例 8-2 Gulf Real Estate Properties 公司

Gulf Real Estate Properties 有限责任公司是佛罗里达西南部的一家房地产公司。企业在广告中称自己是“地产专家”。公司对销售进行监督，搜集有关房屋地点、定价、售价和每套售出中花费天数等信息。如果房屋位于墨西哥湾，则称之为看得见海湾的房屋；如果房屋位于墨西哥湾附近的其他海湾或者高尔夫球场，则称之为看不见海湾的房屋。福罗里达州那不勒斯的 Multiple Listing Service 给出了最近售出的40套看得见海湾的房屋和18套看不见海湾的房屋的样本数据，如表8-7所示，价格以千美元计。

管理报告

1. 对40套看得见海湾的房屋，用适当的描述性统计量对三个变量中的每个变量进行汇总。
2. 对18套看不见海湾的房屋，用适当的描述性统计量对三个变量中的每个变量进行汇总。
3. 比较你的汇总结果，讨论有助于房地产代理商了解地产市场的各种统计结果。
4. 对看得见海湾的房屋，求售价的总体均值以及售出中花费天数的总体均值的95%置信区间。解释你的结果。
5. 对看不见海湾的房屋，求售价的总体均值以及售出中花费天数的总体均值的95%置信区间。解释你的结果。
6. 假定分公司的经理要求在40 000美元的边际误差下对看得见海湾的房屋售价的均值进行估计；在15 000美元的边际误差下对看不见海湾的房屋售价的均值进行估计。取置信度为95%，则应选取多大的样本容量？
7. Gulf Real Estate Properties 公司刚刚签订了两份新的订房合同：一套定价589 000美元的看得见海湾的房屋，一套定价285 000美元的看不见海湾的房屋。请估计这两套房子最终的售价以及需花费多少天才能售出？

表 8-7 Gulf Real Estate Properties 公司的销售数据

看得见海湾的房屋			看不见海湾的房屋		
定价	售价	销售时间	定价	售价	销售时间
495.0	475.0	130	217.0	217.0	182
379.0	350.0	71	148.0	135.5	338
529.0	519.0	85	186.5	179.0	122
552.5	534.5	95	239.0	230.0	150
334.9	334.9	119	279.0	267.5	169
550.0	505.0	92	215.0	214.0	58
169.9	165.0	197	279.0	259.0	110
210.0	210.0	56	179.0	176.5	130
975.0	945.0	73	149.9	144.9	149
314.0	314.0	126	235.0	230.0	114
315.0	305.0	88	199.8	192.0	120
885.0	800.0	282	210.0	195.0	61
975.0	975.0	100	226.0	212.0	146
469.0	445.0	56	149.9	146.5	137
329.0	305.0	49	160.0	160.0	281
365.0	330.0	48	322.0	292.5	63
332.0	312.0	88	187.5	179.0	48
520.0	495.0	161	247.0	227.0	52
425.0	405.0	149			
675.0	669.0	142			
409.0	400.0	28			
649.0	649.0	29			
319.0	305.0	140			
425.0	410.0	85			
359.0	340.0	107			
469.0	449.0	72			
895.0	875.0	129			
439.0	430.0	160			
435.0	400.0	206			
235.0	227.0	91			
638.0	618.0	100			
629.0	600.0	97			
329.0	309.0	114			
595.0	555.0	45			
339.0	315.0	150			
215.0	200.0	48			
395.0	375.0	135			
449.0	425.0	53			
499.0	465.0	86			
439.0	428.5	158			

案例 8-3 Metropolitan Research 有限公司

Metropolitan Research 有限公司是一家消费者研究组织，它设计调查用于对消费者所使用的大量产品和服务进行评估。在某一项特别的研究中，Metropolitan 调查消费者对底特律地区某个主要制造商所生产的汽车性能的满意程度。就该制造商生产的一款正常尺寸的小汽车向用户的问卷调查显示，有许多人抱怨车辆初期传动系统不佳。

为了更好地了解传动系统的问题，Metropolitan 以底特律地区一个修理企业提供的实际传动系统维修记录为样本进行研究。以下为50辆汽车的传动系统出现故障时所行驶的实际里程的数据。

85 092	32 609	59 465	77 437	32 534	64 090	32 464	59 902
39 323	89 641	94 219	116 803	92 857	63 436	65 605	85 861
64 342	61 978	67 998	59 817	101 769	95 774	121 352	69 568
74 276	66 998	40 001	72 069	25 066	77 098	69 922	35 662
74 425	67 202	118 444	53 500	79 294	64 544	86 813	116 269
37 831	89 341	73 341	85 288	138 114	53 402	85 586	82 256
77 539	88 798						

管理报告

1. 利用适当的描述性统计量对传动系统失灵的数据进行汇总。

2. 对曾经出现过传动系统失灵的汽车总体，求出现传动系统问题时行驶里程均值的95%置信区间。并对该区间估计作出管理上的解释。

3. 对于一些汽车用户曾发生过的初期传动系统失灵，你的统计结果说明了什么？

4. 如果想要在5 000英里的边际误差下，估计出现传动系统失灵时所行驶里程的均值，则研究者应选取多大的样本容量？取置信度为95%。

5. 为了更全面地对该传动系统失灵作出评价，你还需要收集一些什么信息？

第9章
假设检验

CHAPTER 9

实践中的统计[㊀]

John Morrell 有限公司

俄亥俄州，辛辛那提市

John Morrell 公司于1827年创建于英国，是美国历史上最悠久的连续经营的肉类制造商。它是一家全资独立管理的子公司隶属于弗吉尼亚州的 Smithfield 食品公司。John Morrell 公司为顾客提供一系列经过加工的肉制品和新鲜的猪肉产品，其产品包括以下13个地区品牌：John Morrell、E-Z-Cut、Tobin's First Prize、Dinner Bell、Hunter、Kretschmar、Rath、Rodeo、Shenson、Farmers Hichory Brand、Iowa Quality 和 Peyton's。每种地区品牌都在消费者中拥有很高的品牌认知度和忠诚度。

John Morrell 公司的市场研究部门负责管理公司各类产品的大量信息，并将这些产品与同类品牌的竞争产品进行对比。最近，该部门对公司生产的炖牛肉与其他两家主要竞争对手的类似牛肉制品进行了比较。在这三种产品的对比检验中，样本中的消费者根据口感、外观、香味和整体偏好，对产品进行评价。

公司关注的一个问题是，消费者总体中是否有超过50%的人偏爱 John Morrell 公司生产的炖牛肉食品。令p表示偏爱 John Morrell 产品的总体比率，研究中面临的假设检验是：

$$H_0: p \leq 0.50 \quad H_a: p > 0.50$$

原假设H_0表示，偏爱 Morrell 产品的总体比率小于或等于50%。如果样本数据支持拒绝H_0，从而接受备择假设H_a，则 John Morrell 会得出研究结论：通过三种产品的比较，消费者总体中超过50%以上的人偏爱 John Morrell 公司的产品。

在一项独立的口感检验研究中，来自辛辛那提、密尔沃基和洛杉矶的224名消费者组成一个样本，其中有150人选择 John Morrell 公司生产的炖牛肉为自己最喜爱的食品。利用统计假设检验方法，原假设H_0被拒绝。研究提供统计证据支持H_a，得出的结论是：消费者总体中超过50%以上的人偏爱 John Morrell 公司的产品。

总体比率的点估计计$\bar{p} = 150/224 = 0.67$。因此，样本数据支持食品杂志的广告，认为在三种产品的口味比较中，John Morrell 公司生产的炖牛肉“在竞争中受欢迎程度为2：1”。

本章我们将介绍如何提出假设，如何像 John Morrell 公司那样进行假设检验。通过分析样本数据，我们可以确定是否拒绝一个假设。

在第7章和第8章中，我们说明了如何利用样本对总体参数进行点估计和区间估计。本章，我们将继续讨论在统计推断中如何利用假设检验来确定是否应拒绝关于总体参数值的说法。

在假设检验中，我们首先对总体参数做一个尝试性的假设。该尝试性的假设称为**原假设**（null hypothesis），记作H_0。然后，定义另一个与原假设的内容完全对立的假设，记作H_a，称之为**备择假设**（alternative hypothesis）。假设检验的过程就是根据样本数据对这两个对立的假设H_0和H_a进行检验。

本章说明如何对总体均值和总体比率进行假设检验。我们首先通过举例说明建立原假设和备择假设的方法。

9.1 原假设和备择假设的建立

原假设和备择假设可能并不总是显而易见的，必须谨慎地构造适当的假设，从而使得假设检验的结论能够提供研究者或者决策者所需要的信息[㊁]。在确定如何表述假设时，假设检验具体应用于何种情况非常重要。在所有假设检验的应用中，都包括搜集样本并利用样本结果提供下结论的依据。在确定原假设和备择假设时，关键的问题是考虑搜集样本的目的是什么，我们想要作出怎样的结论。

㊀ 作者感谢 John Morrell 公司市场营销部副总经理 Marty Butler，他为“实践中的统计”提供了本案例。

㊁ 正确表达假设是需要通过实践来学习的。可以预见，在选择合适的假设H_0和H_a时，最初可能会有一些困惑。本节的例子试图对此提供一些应用指南。

在本章的开篇，我们说原假设 H_0 是关于总体参数（比如，总体均值或者总体比率）的一个尝试性的假设。备择假设 H_a 是与原假设对立的一种说法。在有些情形下，更易于首先确定备择假设，然后再确定原假设。而在另一些情形下，则更易于首先确定原假设，然后再确定备择假设。在下面的例子中，我们将说明这些情况。

9.1.1 将研究中的假设作为备择假设

在许多假设检验的应用中，都是试图搜集证据来支持研究中的假设。在这些情形下，通常最好从备择假设开始，然后得到研究者希望支持的结论。考虑某种汽车，在城市行驶时目前达到的燃油效率为 24 英里/加仑。某产品研究小组专门设计了一种新型的燃油喷射系统来提高每加仑燃油的效率。产品研究小组要对这种新型燃油喷射系统进行控制实验，寻找统计依据作出结论：新型燃油喷射系统的燃油效率超过了目前所使用系统的燃油效率。

制造一批这种新型燃油喷射系统，将它们安装在要检测的汽车上，并且满足研究控制的驾驶条件。计算这些汽车每加仑燃油行驶里程的样本均值，并据此在假设检验中判断是否可以得出结论：新型燃油喷射系统的平均效率超过了 24 英里/加仑。令燃油效率的总体均值为 μ，则研究中的假设 $\mu>24$ 是备择假设。由于目前燃油喷射系统的平均效率为 24 英里/加仑，所以我们尝试性地假定：新型燃油喷射系统并不比目前的系统好，即选择 $\mu\leqslant24$ 为原假设。原假设和备择假设分别为

$$H_0: \mu \leqslant 24$$

$$H_a: \mu > 24$$

如果样本支持得出拒绝 H_0 的结论，则可以作出 $H_a: \mu>24$ 的推断。结论从统计上支持研究者，认为新型燃油喷射系统提高了每加仑燃油的平均行驶里程。因而，可以考虑将此种新型燃油喷射系统投入生产。然而，如果抽样结果得出不能拒绝 H_0 的结论，则研究者不能断定新型燃油喷射系统是否一定优于目前的系统。因此，若从省油角度来看，没有依据表明生产新型燃油喷射系统的汽车是合理的。也许应该进行更深入的研究和检验。[⊖]

企业通过不断开发比已有产品和方法更优的新产品、新方法和新系统等，方能成功地保持竞争力。在接纳一项新事物之前，希望通过研究来判定是否有统计依据支持我们得出新方法确实更好的结论。在这种情形下，通常将研究中的假设表述为备择假设。例如，认为一种新型教学方法优于目前的方法，则备择假设为新方法更好，原假设为新方法不比老方法好。比如，为促销而开展一项新的销售奖励促进计划，则备择假设为新的奖励计划能够促进销售，原假设为新的奖励计划不能促进销售。再比如，为了比已有药物更好地降低血压，研制出了一种新型药物，则备择假设为新型降压药优于已有药物，原假设为新药并不比已有药物存在更好的降压效果。在这些情形下，若拒绝原假设 H_0 则为研究中的假设提供了统计上的支持。在本章以及本书的其余部分，我们还将看到许多研究中假设检验的例子。

9.1.2 将受到挑战的假说作为原假设

当然，不是所有的假设检验都是涉及研究的假设。在下面的讨论中，我们考虑假设检验的另一个应用。我们从一种信念或假定开始，即从有关总体参数值的说法是真实的开始。然后，我们将利用假设检验对这种假定提出怀疑，并确定是否有统计证据支持得出假定不正确的结论。在这种情形下，首先确立原假设是有益的。原假设 H_0 表述了对总体参数值的信念或者假定。在备择假设 H_a 中，认为这种信念或者假说不正确。

作为一个例子，考虑某种软饮料产品的制造商所面对的情况。瓶装软饮料的标签上注明，每瓶为 67.6 盎司。我们认为瓶装软饮料的标签是正确的，即罐装重量的总体均值至少为 67.6 盎司。除非有其他理由，我们通常假设标签上的标注是正确的，但可以对制造商的说法提出质疑。因此，在关于每瓶饮料重量的总体均值的假设检验中，我们会从假设标签是正确的开始，设定原假设为 $\mu\geqslant67.6$ [⊜]。挑战这种假说，意味着标签是不正确的，每瓶分量不足。将这种质疑表述为备择假设 $\mu<67.6$。于是，原假设和备择假设分别为：

$$H_0: \mu \geqslant 67.6$$

$$H_a: \mu < 67.6$$

⊖ 如果样本数据提供了足够的证据表明可以拒绝原假设，那么可以得出结论认为所研究的假设为真。

⊜ 制造商的产品信息通常被假设为真，并将此设为原假设。如果拒绝原假设，则得出该信息不正确的结论。

政府监督部门肩负着保证商标标注有效的义务，他们选取一些瓶装软饮料组成样本，计算罐装重量的样本均值，并据此进行随后的假设检验。如果抽样结果使得我们得到拒绝 H_0的结论，则推断 $H_a:\mu<67.6$ 为真。根据这一统计证据，监督部门可以得出结论：标签不正确，产品分量不足。考虑采取适当的强制措施促使制造商生产的产品达到标签上的标准。但是，如果抽样结果表明不能拒绝 H_0，则不能拒绝“制造商的标签是正确的”这一假设。在这种结论之下，监督部门不需要采取任何行动。

下面我们从制造商的角度重新考虑瓶装软饮料的罐装问题。设计的软饮料的装瓶标准为67.6盎司。制造商不想让产品分量不足，因为这会导致来自消费者或监管部门的投诉。然而，制造商也不想让产品超量，因为灌入过量的软饮料会增加不必要的成本。企业的目标是，调整装瓶操作使得每瓶饮料容量的总体均值恰好为标签标注的67.6盎司。

虽然企业的目标如此，但是生产操作中时常会出现失控。如果发生失控，那么在我们的例子中，将会发生软饮料灌装不足或者灌装过量。无论发生哪一种情形，企业都需要对此了解并加以纠正，重新将瓶装操作调整到设计的67.6盎司。在应用假设检验时，我们依然会从假设生产线运行正常开始，设定原假设为 $\mu=67.6$ 盎司，挑战该假说的备择假设为 $\mu\neq67.6$，即会发生软饮料灌装不足或者灌装过量。在厂商的假设检验中，原假设和备择假设分别为：

$$H_0:\mu = 67.6$$
$$H_a:\mu \neq 67.6$$

假如软饮料制造商利用质量控制方法定期从灌装生产线上抽取一部分瓶装软饮料组成样本，计算每瓶灌入重量的样本均值。如果样本结果使得我们得到拒绝 H_0的结论，则推断 $H_a:\mu\neq67.6$ 为真。我们的结论是：灌装操作运行不正常，需要对生产流程进行调整，以恢复到每瓶的总体均值为67.6盎司的状态。但是，如果样本结果表明不能拒绝 H_0，则不能拒绝“制造商的装瓶操作运行正常”这一假设。这时，不需要采取任何行动，生产操作继续进行。

在关于软饮料制造的例子中，上述两种假设检验的形式说明：研究者或者决策者的观点不同，原假设和备择假设也随之不同。在正确地建立假设时，重要的是，要了解假设检验具体应用于何种情况，从而使得构建的假设能够提供研究者或者决策者所需要的信息。

9.1.3 原假设和备择假设形式的小结

本章中的假设检验涉及两个总体参数：总体均值和总体比率。根据不同情况，总体参数的假设检验采取以下三种形式之一：其中两个在原假设用不等号（≥或≤）；第三个在原假设用等号。对于总体均值的假设检验，我们令 μ_0 代表假定值并且必须采用以下三种形式[⊖]之一进行假设检验。

$$H_0:\mu \geq \mu_0 \qquad H_0:\mu \leq \mu_0 \qquad H_0:\mu = \mu_0$$
$$H_a:\mu < \mu_0 \qquad H_a:\mu > \mu_0 \qquad H_a:\mu \neq \mu_0$$

前面两种形式称为单侧检验；第三种形式称为双侧检验。

许多情况下，关于 H_0和 H_a的选择并非显而易见，必须通过判断来选择。当然，正如前面的形式所显示的那样，表达式中的等号部分（不论是≥，≤还是=）总是出现在原假设中。在选择 H_0和 H_a的适当形式时，记住将检验试图建立的结果设为备择假设，因此，询问使用者是否正在寻找支持 $\mu<\mu_0$，$\mu>\mu_0$或 $\mu\neq\mu_0$的证据，有助于我们确定 H_a。下面设计的习题，让我们练习如何在总体均值的假设检验中选取假设的适当形式。

练习

☆2. 一名汽车销售经理正在考虑采取一种新的奖励计划以提高销售量。目前，销售量的均值为每月14辆汽车。经理通过调研想知道新的奖励计划能否增加销售量。为了收集有关新计划的数据，允许被选入样本的销售人员在为期一个月的时间内在销售中采用新的奖励计划。

⊖ 这里给出了假设 H_0和 H_a的三种可能的形式，注意等号总是出现在原假设 H_0中。

a. 在这种情况下，建立最为合适的原假设和备择假设。

b. 若不能拒绝 H_0，请对所做的结论进行评述。

c. 若能拒绝 H_0，请对所做的结论进行评述。

4. 由于改变生产方法所需的时间和成本都很大，所以在一种新方法实施之前，制造负责人必须使管理人员确信，推荐的制造方法能降低成本。目前生产方法的平均成本是每小时 220 美元。一项调研工作是在样本生产期间，测量新方法的成本。

a. 在该项研究中，建立恰当的原假设和备择假设。

b. 当 H_0 不能被拒绝时，对所做的结论进行评述。

c. 当 H_0 能被拒绝时，对所做的结论进行评述。

9.2 第一类错误和第二类错误

原假设和备择假设是关于总体的两种不同的观点。要么原假设 H_0 为真，要么备择假设 H_a 为真，但是两者不可能同时为真。理想的假设检验方法应该是：当 H_0 为真的时候接受 H_0；当 H_a 为真的时候拒绝 H_0。但是，由于假设检验是基于样本信息得到的，不可能作出的结论总是正确的，所以我们必须考虑发生误差的可能性。表 9-1 列示了假设检验中可能发生的两类错误。

表 9-1 的第一行说明，当作出接受 H_0 的结论时可能发生的情况。这时，如果 H_0 为真，则该结论正确，如果 H_a 为真，那么发生了**第二类错误**（Type II error），即当 H_0 为假时我们却接受了 H_0。表 9-1 的第二行说明，当作出拒绝 H_0 结论时可能发生的情况。这时，如果 H_0 为真，那么发生了**第一类错误**（Type I error），即当 H_0 为真时我们却拒绝了 H_0。显然，如果 H_a 为真，则拒绝 H_0 是正确的。

表 9-1 假设检验中的正确与错误结论

		总体情况	
		H_0 是真的	H_a 是真的
结论	接受 H_0	结论正确	第二类错误
	拒绝 H_0	第一类错误	结论正确

回忆在第 9.1 节讨论的假设检验的例子。一个汽车生产研究小组开发了一种新型燃油喷射系统，该系统能提高汽车燃油的效率。目前所用系统下，汽车燃油的效率为 24 英里/加仑，建立如下形式的假设检验

$$H_0: \mu \leqslant 24$$

$$H_a: \mu > 24$$

备择假设 $H_a: \mu > 24$ 表明，研究者正致力于寻找样本证据支持结论：新型燃油喷射系统能使汽车燃油效率的总体均值超过 24 英里/加仑。

在这个应用中，第一类错误是指 H_0 为真时却拒绝了 H_0，即当研究者认为新型燃油喷射系统提高了燃油效率（$\mu > 24$），而实际情况却是新系统并不比目前所使用的系统好。反之，第二类错误是指 H_0 为假时却接受了 H_0，即当研究者在认为新型燃油喷射系统的效率并不比目前的好（$\mu \leqslant 24$），而实际情况却是新系统提高了每加仑燃料汽车的行驶里程。

对于每加仑燃料行驶里程的假设检验，原假设是 $H_0: \mu \leqslant 24$。假定，原假设为真是以等式形式出现，即 $\mu = 24$。当作为一个等式的原假设为真时，犯第一类错误的概率被称为检验的**显著性水平**（level of signifcance）。于是，对于每加仑燃料汽车行驶里程的假设检验而言，显著性水平是当 $\mu = 24$ 时拒绝 $H_0: \mu \leqslant 24$ 的概率。出于这个概念的重要性，下面我们重新叙述显著性水平的定义。

显著性水平

当作为一个等式的原假设为真时，犯第一类错误的概率称为检验的显著性水平。

用希腊字母 α（阿尔法）表示显著性水平，一般取 α 为 0.05 和 0.01。

在实践中，由进行假设检验的人设定显著性水平。通过选择 α，控制了犯第一类错误的概率。如果犯第一类错误的成本很高，则选择小的 α 值。如果犯第一类错误的成本不高，则通常选择较大的 α 值。应用中，一般将只控制第一类错误的假设检验称为显著性检验。许多假设检验的应用都属于这一种类型。

在假设检验的大多数应用中，虽然对发生第一类错误的概率进行了控制，但通常并不对发生第二类错误的概率加以控制。因此，如果我们决定接受 H_0，我们并不能确定该决策有多大的可信度。由于显著性检验中第二类错

误的发生具有不确定性，所以通常统计学家建议我们在叙述中采用“不能拒绝 H_0”而不采用“接受 H_0”这种说法。采用“不能拒绝 H_0”这种说法意味着我们对判断或行动持保留意见。实际上，不直接接受 H_0，使统计学家避免了发生第二类错误的风险。不论何时，只要未对发生第二类错误的概率加以确定或控制，我们就不能作出接受 H_0的结论。这种情形下，只能得出两种可能的结论：不能拒绝 H_0或拒绝 H_0。[⊖]

虽然在假设检验中通常没有控制第二类错误，但是第二类错误确实是可以控制的。在第9.7和第9.8节中，我们将介绍确定和控制第二类错误发生概率的方法。如果能够对第二类错误建立适当的控制，那么基于“接受 H_0”的结论所采取的措施就是恰当的。

注释和评论

美国 George Mason 大学的经济学教授兼联合专栏作家沃尔特·威廉姆斯（Walter Williams）指出，在决策中第一类错误或第二类错误发生的可能性经常存在（*The Cincinnati Enquirer*，2005年8月4日）。他认为，美国食品和药品管理局（FDA）在药品准入程序中具有犯这些错误的风险。对某种药物，FDA 必须确定是否准许其用于诊疗。第一类错误意味着，FDA 准许一种并非安全有效的药物投入使用。而第二类错误意味着，FDA 不允许一种安全有效的药物用于诊疗。无论作出何种决策，发生重大损失的可能性都是不能忽视的。

练习

6. 在容量为3夸脱的橙汁容器上，标签标明橙汁脂肪含量的均值不超过1克。对标签上的说明进行假设检验，回答下列问题。
   a. 建立适当的原假设和备择假设。
   b. 这种情况下，第一类错误是什么？发生这类错误的后果是什么？
   c. 这种情况下，第二类错误是什么？发生这类错误的后果是什么？

8. 如果假设检验支持我们作出新的生产方法能够降低每小时操作成本的结论，则将采用这种新的生产方法。
   a. 如果目前生产方法的平均成本为每小时220美元，建立合适的原假设和备择假设。
   b. 这种情况下，第一类错误是什么？发生这类错误的后果是什么？
   c. 这种情况下，第二类错误是什么？发生这类错误的后果是什么？

9.3 总体均值的检验：σ已知情形

在第8章中，我们曾说如果在应用中可以根据历史数据或者其他信息在抽样前得到总体标准差的一个好的估计，则可看作 σ 是已知的。在这种情形下，出于应用的目的，总体标准差可以被认为是已知的。本节我们说明，在 σ 已知的情形下如何对总体均值进行假设检验。

在样本恰好选自服从正态分布的总体时，本节所介绍的方法是精确成立的。当没有理由可以假定总体服从正态分布的情形下，只有在样本容量足够大的时候，这些方法才是有效的。在本节末，我们将就总体分布和样本容量给出在实际应用中的一些建议。

9.3.1 单侧检验

总体均值的单侧检验（one-tailed test）有以下两种形式。

下侧检验	上侧检验
$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$
$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$

⊖ 如果样本数据与原假设 H_0一致，则我们得出“不能拒绝 H_0”的结论。因为接受 H_0使我们有发生第二类错误的风险，所以更倾向于“不能拒绝 H_0”而不是“接受 H_0”。

我们考虑一个下侧检验的例子。

美国联邦贸易委员会（FTC）定期设计统计调查，用以检验制造商的产品说明。例如，大号听装 Hilltop 咖啡的标签上标明装有 3 磅咖啡。FTC 知道 Hilltop 的生产线不可能精确地在每罐中放入 3 磅咖啡，甚至无法保证所有听装咖啡重量的总体均值为 3 磅/听。当然，只要听装总体重量的均值至少为 3 磅/听，消费者的权益将得到保障。于是，FTC 把大号听装咖啡标签上的信息理解为 Hilltop 的承诺：听装咖啡重量的总体均值为 3 磅/听。我们将说明，FTC 如何通过下侧检验来验证 Hilltop 的承诺。

第一步是为检验提出原假设和备择假设。如果罐内装入咖啡重量的总体均值至少为 3 磅/听，那么 Hilltop 遵守了产品的承诺。这构成了检验的原假设。当然，如果罐内装入咖啡重量的总体均值少于 3 磅/听，那么 Hilltop 没有遵守对产品的承诺，这构成检验的备择假设。令 μ 代表听装咖啡重量的总体均值，得到如下形式的原假设和对立假设：

$$H_0: \mu \geqslant 3 \qquad H_a: \mu < 3$$

注意，总体均值的假定值 $\mu_0 = 3$。

如果样本数据表明，不能拒绝 H_0，则统计证据不支持我们作出生产商违背了标签上承诺的结论，从而不必对 Hilltop 采取任何措施。但是，如果样本数据表明能拒绝 H_0，则我们的结论是：备择假设 $H_a: \mu < 3$ 为真。这种情形下，可以作出听装咖啡重量不足的结论，并且有充足的理由投诉 Hilltop 违背了标签上的承诺。

假定选取 36 听咖啡组成一个随机样本，并且计算样本均值 $\bar{x}$ 作为总体均值 μ 的估计值。如果样本均值 $\bar{x}$ 小于 3 磅，则样本结果对原假设提出了质疑。我们想知道的是：当样本均值 $\bar{x}$ 比 3 磅少多少的时候，我们才能断言差异明显并且甘愿冒着犯第一类错误的风险（错误地）控告 Hilltop 公司违背了标签上的承诺。这个问题中，一个关键的因素是决策者所选取的显著性水平。

正如上一节所提到的那样，用 α 表示显著性水平，它是当作为一个等式的原假设为真时，拒绝 H_0 而发生第一类错误的概率。决策者必须事先指定一个显著性水平。如果发生第一类错误的成本很高，则应当选取较小的值做为显著性水平。如果发生第一类错误的成本不高，则可以适当选取较大的值为显著性水平。在 Hilltop 咖啡的研究中，FTC 检验程序的负责人作出如下说明："如果公司达到了其重量规格 $\mu = 3$ 的要求，则我们不会对 Hilltop 采取任何措施。但是，我们会以 1% 的可能性犯这类错误。"从该负责人的说明中可知，我们设定该假设检验的显著性水平 $\alpha = 0.01$。于是，我们必须设计一个假设检验，使得当 $\mu = 3$ 时犯第一类错误的概率为 0.01。

在 Hilltop 咖啡问题的研究中，通过确定原假设、备择假设和设定显著性水平，我们已经完成了进行每个假设检验时所需的前两步。现在我们准备完成假设检验的第三步：收集样本数据和计算检验统计量的值。

检验统计量（test statistic）在 Hilltop 咖啡问题的研究中，FTC 前期的检验表明，可以假定总体标准差已知，为 $\sigma = 0.18$。另外，前期检验还表明，可以假定听装咖啡重量的总体服从正态分布。从第 7 章抽样分布的研究中我们知道，如果进行抽样的总体服从正态分布，则 $\bar{x}$ 的抽样分布也服从正态分布。因此，在 Hilltop 咖啡的研究中，$\bar{x}$ 的抽样分布服从正态分布，已知 $\sigma = 0.18$，样本容量 $n = 36$。图 9-1 给出当作为一个等式的原假设为真（即 $\mu = \mu_0 = 3$ ㊀）时 $\bar{x}$ 的抽样分布。注意：$\bar{x}$ 的标准误差，$\sigma_{\bar{x}} = \sigma / \sqrt{n} = 0.18 / \sqrt{36} = 0.03$ ㊁。

由于 $\bar{x}$ 的抽样分布服从正态分布，则

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - 3}{0.03}$$

的抽样分布是标准正态分布。$z = -1$ 表明 $\bar{x}$ 的值位于比均值的假定值小 1 个标准差的位置，$z = -2$ 表明 $\bar{x}$ 的值位于比均值的假定值小 2 个标准差的位置，…，依次类推。我们可以通过标准正态概率分布表得到对应于任何 z 值的下侧概率。例如，$z = -3$ 的下侧面积为 0.001 3。从而，

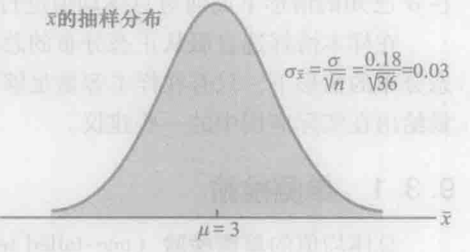


图 9-1 原假设以等式（$\mu = 3$）形式成立时，在 Hilltop 咖啡问题的研究中，$\bar{x}$ 的抽样分布

㊀ 在构造假设检验的抽样分布时，假定 H_0 是以等式形式成立的。

㊁ $\bar{x}$ 的标准误差就是 $\bar{x}$ 的抽样分布的标准差。

所得 z 值小于均值3个或3个以上标准差的概率为0.0013。因此，$\bar{x}$ 的值比总体均值的假定值 $\mu_0=3$ 小3个或3个以上标准差的概率为0.0013。如果原假设为真，这个结果是不大可能发生的。

在 σ 已知的情形下对总体均值进行假设检验，我们用标准正态随机变量 z 作为**检验统计量**（test statistic）来确定 $\bar{x}$ 是否偏离假定值 μ 足够远，从而有理由拒绝原假设。令 $\sigma_{\bar{x}}=\sigma/\sqrt{n}$，检验统计量如下。

总体均值假设检验的检验统计量：σ 已知的情形

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \tag{9-1}$$

下侧检验的关键问题在于：检验统计量 z 的值必须达到多小的时候，我们才能选择拒绝原假设？有两种方法可以解决这个问题：p-值法和临界值法。

***p*-值法** p-值法利用检验统计量 z 的值计算一个被称为 ***p*-值**（p-value）的概率。

***p*-值**

p-值是一个概率值，它度量样本所提供的证据对原假设的支持程度。p-值越小说明反对原假设的证据越多。[⊖]

p-值用于确定是否拒绝原假设。我们看看如何计算和使用 p-值。利用检验统计量计算 p-值。用于计算 p-值的方法依赖于检验是下侧检验、上侧检验还是双侧检验。对于下侧检验，p-值是检验统计量小于或等于样本所给出的检验统计量的值的概率。从而，在 σ 已知的情形下为了计算下侧检验的 p-值，我们必须得到标准正态曲线下在检验统计量的值左边部分的面积。在计算出 p-值以后，我们必须决定它是否小到足以拒绝原假设；正如我们将要说明的那样，这需要将计算出的 p-值与显著性水平进行比较。

我们现在计算 Hilltop 咖啡问题中下侧检验的 p-值。假定选取36听 Hilltop 咖啡组成一个样本，样本均值 $\bar{x}=2.92$ 磅。$\bar{x}=2.92$ 是否小到足以拒绝 H_0？因为这是一个下侧检验，p-值是标准正态曲线下在检验统计量的值左边部分的面积。利用 $\bar{x}=2.92$，$\sigma=0.18$ 和 $n=36$，我们计算检验统计量 z 的值为

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{2.92-3}{0.18/\sqrt{36}}=-2.67$$

从而，p-值为检验统计量 z 小于或等于 -2.67 的概率（标准正态曲线下在检验统计量的值下侧部分的面积）。

利用标准正态概率表，我们查得 $z=-2.67$ 下侧的面积为0.0038。图9-2给出了与 $\bar{x}=2.92$ 相对应的检验统计量 $z=-2.67$ 和 p-值 $=0.0038$。p-值表明，当从 $\mu=3$ 的总体进行抽样时，所得到的样本均值为 $\bar{x}=2.92$（相应的检验统计量为 -2.67）或者更小的概率是很小的。这个 p-值没有支持原假设，但是它小到足以能够使我们拒绝 H_0 了吗？答案依赖于检验的显著性水平。

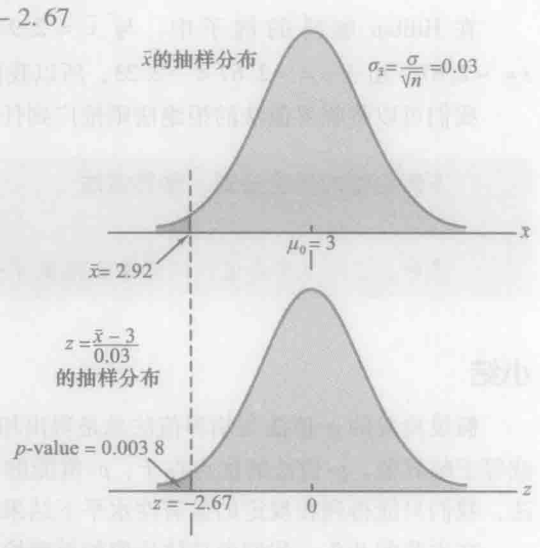


图9-2 Hilltop 咖啡的研究中，当 $\bar{x}=2.92$ 和 $z=-2.67$ 时，求得的 p-值

正如前面所说，FTC 检验的项目负责人选取0.01作为显著性水平。选择 $\alpha=0.01$ 意味着：在原假设以等式形式（即 $\mu_0=3$）成立时，负责人愿意容许以0.01的概率拒绝原假设。在 Hilltop 问题的研究中，由36听咖啡组成一个样本，所得 p-值 $=0.0038$，这意味着：当作为一个等式的原假设为真时，得到 $\bar{x}=2.92$ 或者更小 $\bar{x}$ 的值的概率为0.0038。由于0.0038小于或者等于 $\alpha=0.01$，故我

⊖ 小的 p-值表明，检验统计量的值通常不能得出假设 H_0 为真的结论。

们拒绝 H_0，从而在 0.01 的显著性水平下我们发现有足够的统计证据拒绝原假设。

我们现在可以陈述利用 p-值法确定是否可以拒绝原假设的基本规则。对于给定的显著性水平 α，p-值法的拒绝法则如下。

p-值法的拒绝法则

如果 p-值 $\leqslant \alpha$，则拒绝 H_0

在 Hilltop 咖啡的检验中，p-值 = 0.003 8，于是拒绝原假设。是否拒绝原假设需要将 p-值与 FTC 负责人选取的显著性水平进行比较。实测的 p-值 = 0.003 8 意味着对于任何 $\alpha \geqslant 0.0038$ 的情形，我们都将拒绝 H_0。因此，p-值也被称为实际显著性水平。

不同的决策者可能对发生第一类错误的成本有不同的看法，从而选择不同的显著性水平。将 p-值做为假设检验的一部分，决策者可以将报告的 p-值与自己的显著性水平进行比较，从而可对是否拒绝 H_0 作出各自不同的决定。

临界值法 临界值法要求我们首先确定被称为临界值的检验统计量的值。对于下侧检验，**临界值**（critical value）是确定检验统计量的值是否小到足以拒绝原假设的一个基准。在检验统计量的抽样分布中，与下侧面积 α（显著性水平）相对应的值是检验统计量的临界值。换句话说，临界值是使得我们拒绝原假设的检验统计量的最大值。我们回到 Hilltop 咖啡的例子中，看看如何使用临界值法。

在 σ 已知的情形下，检验统计量 z 的抽样分布是标准正态分布。从而，临界值是标准正态概率分布中与下侧面积 $\alpha = 0.01$ 相对应的检验统计量的值。利用标准正态概率分布表，我们发现 $z = -2.23$ 的下侧面积等于 0.01（见图9-3）。从而，如果由样本所得到的检验统计量的值小于或等于 −2.23，则相应的 p-值将小于或等于 0.01。在这种情况下，我们将拒绝原假设。从而，在 Hilltop 咖啡的研究中，在 0.01 的显著性水平下临界值的拒绝法则是

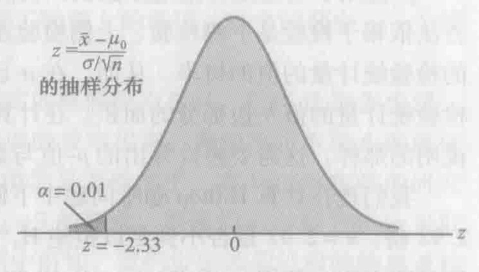


图 9-3 Hilltop 咖啡的假设检验中的临界值是 −2.23

$$\text{如果 } z \leqslant -2.23, \text{则拒绝 } H_0$$

在 Hilltop 咖啡的例子中，与 $\bar{x} = 2.92$ 相对应的检验统计量 $z = -2.67$。由于 $z = -2.67 < -2.23$，所以我们拒绝 H_0 并且得出结论认为 Hilltop 咖啡的分量不足。

我们可以将临界值法的拒绝法则推广到任意的显著性水平。下侧检验的拒绝法则如下。

下侧检验的拒绝法则：临界值法

$$\text{如果 } z \leqslant -z_\alpha, \text{则拒绝 } H_0$$

式中，$-z_\alpha$ 为临界值，即标准正态概率分布下侧的面积为 α 时对应的 z 值。

小结

假设检验的 p-值法与临界值法总是得出相同的拒绝结论，即每当 p-值小于等于 α 时，检验统计量的值将小于或等于临界值。p-值法的优点在于，p-值能够告诉我们结果有多么显著（实际显著性水平）。而如果使用临界值法，我们只能得到在规定的显著性水平下结果是否显著。

在本节的开头，我们说总体均值的单侧检验采用以下两种形式之一：

下侧检验	上侧检验
$H_0: \mu \geqslant \mu_0$	$H_0: \mu \leqslant \mu_0$
$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$

我们通过对 Hilltop 咖啡问题的研究说明了如何进行下侧检验。我们可以用相同的思路进行上侧检验。仍然通

过式（9-1）计算检验统计量。但是，对于上侧检验，p-值是检验统计量大于或等于样本所给出的检验统计量的值的概率。从而，为了在 σ 已知的情形下计算上侧检验的 p-值，我们必须利用标准正态分布求出 z 大于或等于检验统计量值的概率。采用临界值法时，如果检验统计量的值大于或等于临界值 z_α，则我们将拒绝原假设；换句话说，如果 $z \geqslant z_\alpha$，则拒绝 H_0。

单侧假设检验中 p-值的计算步骤总结如下。

计算单侧检验的 p-值

1. 用式（9-1）计算检验统计量的值。
2. 下侧检验：根据标准正态分布，计算 z 小于或者等于检验统计量的值的概率（下侧面积）。
3. 上侧检验：根据标准正态分布，计算 z 大于或者等于检验统计量的值的概率（上侧面积）。

9.3.2 双侧检验

在假设检验中，关于总体均值的**双侧检验**（two-tailed test）的一般形式如下：

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$

在这一小节，我们将介绍如何在 σ 已知情形下对总体均值进行双侧检验。我们以 Maxflight 有限公司的假设检验为例。

制定了一些标准，高尔夫设备制造商如果想让他们的产品用于美国高尔夫球联合会（USGA）的赛事，则必须达到 USGA 的一套标准。Maxflight 有限公司最近采用一种高技术制造工艺生产高尔夫球，其平均发球距离达到 295 码。然而，这种制造工艺有时发生调控失常，导致所生产的高尔夫球的平均发球距离不是 295 码。当平均发球距离低于 295 码时，公司担心由于高尔夫球的平均发球距离没有达到广告中宣传的那么远而使销售量减少。当球的平均发球距离超过 295 码时，Maxflight 的高尔夫球将因为超过 USGA 制定的击出和滚动总距离标准而被 USGA 拒绝。

Maxflight 的质量控制程序定期选择 50 个高尔夫球组成样本来监控制造工艺过程。对每个样本进行假设检验，确定制造工艺是否失控。我们建立原假设和备择假设。我们从假定制造工艺仍然正常运行开始，即所生产的高尔夫球的平均击球距离为 295 码。这个假定构成原假设。备择假设是平均距离不等于 295 码。令假设值 $\mu_0 = 295$，则 Maxflight 假设检验的原假设和备择假设如下：

$$H_0: \mu = 295$$

$$H_a: \mu \neq 295$$

如果样本均值明显小于 295 码或明显大于 295 码，则我们拒绝 H_0。此时，采取的矫正措施是调整制造工艺。另一方面，如果样本均值 $\bar{x}$ 没有明显地偏离假定的均值 $\mu_0 = 295$，将不会拒绝 H_0，也不会采取任何矫正措施去调整制造流程。

质量控制小组选择 $\alpha = 0.05$ 做为检验的显著性水平。在制造流程处于正常运行时，根据前期检验的数据，可以假定已知总体标准差为 $\sigma = 12$。样本容量 $n = 50$ 时，$\bar{x}$ 的标准误差为

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{50}} = 1.7$$

由于样本容量较大，根据中心极限定理可知，$\bar{x}$ 的抽样分布近似服从正态分布（见第 7 章）。图 9-4 是当假定总体均值 $\mu_0 = 295$ 时，Maxflight 假设检验中 $\bar{x}$ 的抽样分布。

假定抽取 50 个高尔夫球组成一个样本，样本均值 $\bar{x} = 297.6$ 码。样本均值支持作出总体均值大于 295 码的结论。$\bar{x}$ 的值超过 295 码是否足够大，使得我们能够在 0.05 的显著性水平下拒绝 H_0。在上一节中，我们介绍了两种方法来回答这个问题：p-值法和临界值法。

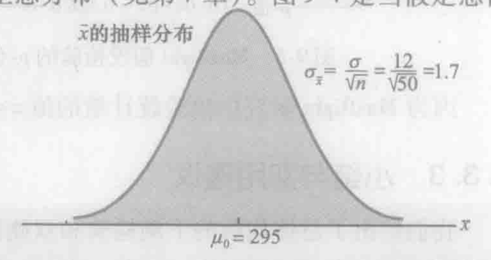


图 9-4 Maxflight 假设检验中 $\bar{x}$ 的抽样分布

p-值法 p-值是一个用于确定是否能够拒绝原假设的概率值。对于双侧检验，若检验统计量的值位于抽样分布的两侧尾部，则支

持拒绝原假设。对于双侧检验，p-值是一个概率值，用来衡量检验统计量与根据样本计算得出的检验统计量值存在明显差异的程度。下面，我们看看在 Maxflight 假设检验中如何计算 p-值。

我们首先计算检验统计量的值。在 σ 已知的情形下，检验统计量 z 是一个标准正态随机变量。利用式（9-1）和 $\bar{x}=297.6$，检验统计量的值为

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{297.6-295}{12/\sqrt{50}}=1.53$$

为了计算 p-值，现在必须计算检验统计量的值与 $z=1.53$ 不同的概率。显然，$z\geqslant1.53$ 的任何值都与 $z=1.53$ 不同；但是，因为这是一个双侧检验，所以 $z\leqslant-1.53$ 的任何值也都与 $z=-1.53$ 不同。在图 9-5 中可见，这个例子中双侧检验的 p-值为 $P(z\leqslant-1.53)+P(z\geqslant1.53)$。因为正态曲线是对称的，可以先计算标准正态曲线下 $z=1.53$ 右侧区域的面积，然后再乘以 2 即得到 p-值。查标准正态分布表可得，$P(z<1.53)=0.9370$。从而，上侧面积 $P(z\geqslant1.53)=1.0000-0.9370=0.0630$，再乘以 2，我们得到 Maxflight 双侧检验的 p-值 $=2\times0.0630=0.1260$。

接下来，我们将 p-值与显著性水平进行比较，看是否应该拒绝原假设。若取显著性水平 $\alpha=0.05$，由于 p-值 $=0.1260>0.05$，故不能拒绝 H_0。因为原假设没有被拒绝，所以不必调整 Maxflight 的制造流程。

与单侧检验中 p-值的计算相比，双侧检验中 p-值的计算看上去似乎略为复杂。双侧假设检验 p-值的计算可简化为如下三个步骤：

双侧检验 p-值的计算

1. 利用式（9-1）计算检验统计量的值。
2. 如果检验统计量的值位于上侧（$z>0$），计算 z 大于或者等于检验统计量值的概率（上侧面积）。如果检验统计量的值位于下侧（$z<0$），计算 z 小于或者等于检验统计量值的概率（下侧面积）。
3. 将步骤 2 所计算出的概率（面积）乘以 2 即求得 p-值。

临界值法 在结束这一节之前，我们看看如何将检验统计量 z 的值和临界值进行比较，从而对一个双侧检验作出假设检验决策。由图 9-6 可见，检验的临界值位于标准正态分布的上侧尾端和下侧尾端。取显著性水平 $\alpha=0.05$ 时，每侧尾端临界值所对应的面积为 $\alpha/2=0.05/2=0.025$。查标准正态概率表可知，检验统计量的临界值 $-z_{0.025}=-1.96$ 和 $z_{0.025}=1.96$。从而，利用临界值法，双侧检验的拒绝法则是

如果 $z\leqslant-1.96$ 或者 $z\geqslant1.96$，则拒绝 H_0

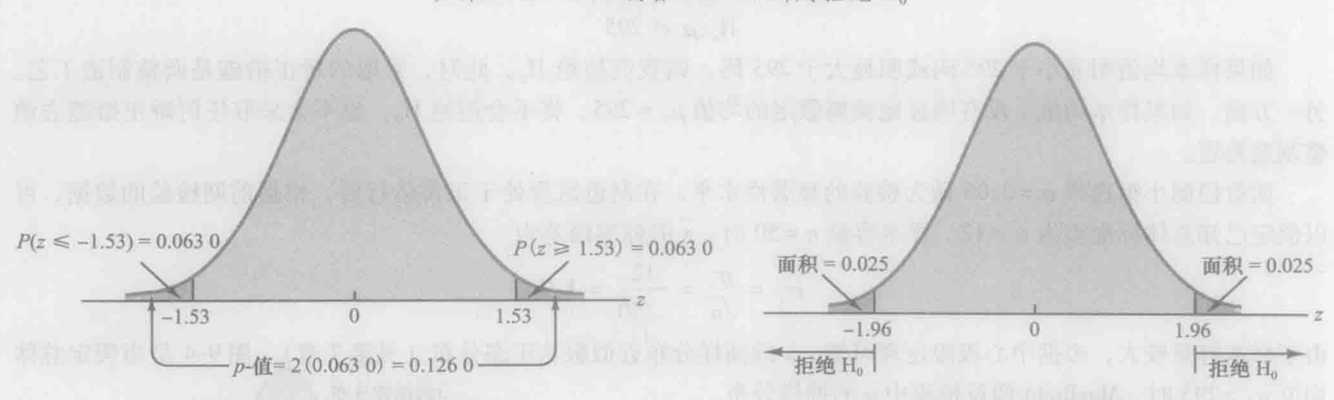


图 9-5 Maxflight 假设检验的 p-值　　图 9-6 Maxflight 假设检验的临界值

因为 Maxflight 研究中检验统计量的值 $z=1.53$，在 0.05 的显著性水平下统计证据不允许我们拒绝原假设。

9.3.3 小结与实用建议

我们给出了总体均值的下侧检验和双侧检验的例子。根据这些例子，我们现在可以将 σ 已知的情形下总体均值的假设检验方法汇总在表 9-2 中。注意，其中 μ_0 是总体均值的假设值。

本节两个例子中假设检验的步骤对于任何假设检验都是适用的。

表 9-2 总体均值假设检验的小结：σ 已知的情形

	下侧检验	上侧检验	双侧检验
假设	$H_0:\mu\geqslant\mu_0$ $H_a:\mu<\mu_0$	$H_0:\mu\leqslant\mu_0$ $H_a:\mu>\mu_0$	$H_0:\mu=\mu_0$ $H_a:\mu\neq\mu_0$
检验统计量	$z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
拒绝法则：p-值法	如果 p-值$\leqslant\alpha$，则拒绝 H_0	如果 p-值$\leqslant\alpha$，则拒绝 H_0	如果 p-值$\leqslant\alpha$，则拒绝 H_0
拒绝法则：临界值法	如果 $z\leqslant -z_\alpha$，则拒绝 H_0	如果 $z\geqslant z_\alpha$，则拒绝 H_0	如果 $z\leqslant -z_{\alpha/2}$ 或者 $z\geqslant z_{\alpha/2}$，则拒绝 H_0

假设检验的步骤

步骤　提出原假设和备择假设。

步骤　指定检验中的显著性水平。

步骤　收集样本数据并计算检验统计量的值。

***p*-值法**

步骤　利用检验统计量的值计算 p-值。

步骤　如果 p-值$\leqslant\alpha$，则拒绝 H_0。

步骤　在应用中解读统计结论。

临界值方法

步骤　利用显著性水平确定临界值以及拒绝法则。

步骤　利用检验统计量的值以及拒绝法则确定是否拒绝 H_0。

步骤　在应用中解读统计结论。

对于实际应用中假设检验的样本容量问题，与我们在第 8 章区间估计中有关样本容量所给出的建议非常类似。在绝大多数应用中，当使用本节介绍的假设检验方法时，样本容量 $n\geqslant30$ 就足够了。在样本容量小于 30 的情形下，则要特别关注抽样的总体分布。如果总体是正态分布的，则我们所描述的假设检验方法是精确的，从而适用于任何样本容量。如果总体不是正态分布但至少大致上是对称的，则可以预期当样本容量 $n\geqslant15$ 时仍能得到可信的结果。

9.3.4 区间估计与假设检验的关系

在第 8 章中我们介绍了如何估计总体均值的置信区间。在 σ 已知的情形下，总体均值的 $100(1-\alpha)\%$ 置信区间估计为

$$\bar{x}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

在本章中，总体均值的双侧假设检验中原假设和备择假设的形式如下：

$$H_0:\mu=\mu_0$$
$$H_a:\mu\neq\mu_0$$

式中，μ_0 为总体均值的假设值。

假定我们沿用第 8 章所描述的方法构造总体均值的 $100(1-\alpha)\%$ 置信区间。在得到的全部置信区间中，有 $100(1-\alpha)\%$ 将包含总体均值，有 $100\alpha\%$ 不包含总体均值。于是，每当置信区间不包含 μ_0 时，如果我们拒绝 H_0，那么我们在原假设为真（$\mu=\mu_0$）时以概率 α 拒绝了原假设。显著性水平是当原假设为真时拒绝原假设的概率。

因此，构造一个$100(1-\alpha)\%$置信区间，并且当置信区间不包含μ_0时拒绝H_0，等价于在显著性水平α下进行双侧假设检验。利用置信区间进行双侧假设检验的方法概括如下。

用置信区间的方法进行如下形式的假设检验

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

1. 从总体中抽取一个简单随机样本，并利用样本均值$\bar{x}$建立总体均值μ的置信区间。

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

2. 如果置信区间包含假设值μ_0，则不能拒绝H_0。否则，拒绝H_0。[⊖]

下面，我们以 MaxFlight 的问题为例，说明如何运用置信区间方法进行假设检验。MaxFlight 的假设检验具有以下形式。

$$H_0: \mu = 295$$
$$H_a: \mu \neq 295$$

假设检验中取显著性水平$\alpha = 0.05$，我们抽取 50 只高尔夫球并计算发球距离的样本均值$\bar{x} = 297.6$码，已知总体标准差$\sigma = 12$码。利用这些结果及$z_{0.025} = 1.96$，得到总体均值的95%置信区间估计为

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$297.6 \pm 1.96 \frac{12}{\sqrt{50}}$$

$$297.6 \pm 3.3$$

或者

$$(294.3, 300.9)$$

根据这一结果，质量控制管理人员可在95%的置信度下得出结论：高尔夫球发球距离的总体均值介于294.3～300.9码。由于总体均值的假设值（$\mu_0 = 295$）落在该区间中，所以假设检验的结论为：不能拒绝原假设H_0: $\mu = 295$。

注意，虽然上述讨论以及例子都是关于总体均值的双侧假设检验，但是，置信区间与双侧假设检验之间的这种对应关系，对其他总体参数也是成立的。这种关系还可以扩展到总体参数的单侧检验中，只不过这时要求建立的是单侧置信区间而已。单侧置信区间在应用中较少遇到。

注释和评论

我们已经说明了如何利用p-值。p值越小，则拒绝H_0的证据越多，从而支持H_a的证据越多。下面是统计学家给出的一些用于解读p-值时的指导意见。

- p-值小于0.01——强有力的证据断定H_a为真。
- p-值介于0.01～0.05——有力证据断定H_a为真。
- p-值介于0.05～0.10——弱的证据断定H_a为真。
- p-值大于0.10——没有足够的证据断定H_a为真。

⊖ 在置信区间方法的拒绝法则中，如果μ_0恰好位于$100(1-\alpha)\%$置信区间的端点则拒绝H_0；在p-值法的拒绝法则中，如果p-值$\leqslant \alpha$则拒绝H_0。在这方面，二者是一致的。

对于双侧假设检验，如果置信区间不包括μ_0，则拒绝原假设。

练 习

提示：下边的习题中有一些要求用 p-值法，另一些要求使用临界值法。这两种方法将得到相同的假设检验结论。我们对这两种方法都提供了习题供你进行练习。下一节以及其余各章中，我们更偏好使用 p-值法。你可以根据个人偏好选择其中任何一种方法。

方法

☆10. 考虑如下假设检验

$$H_0: \mu \leqslant 25$$
$$H_a: \mu > 25$$

一个容量为40的样本中，样本均值为26.4。总体标准差为6。

a. 计算检验统计量的值。

b. p-值是多少？

c. $\alpha = 0.01$ 时，你的结论是什么？

d. 临界值法中的拒绝法则是什么？你的结论是什么？

12. 考虑下面的假设检验

$$H_0: \mu \geqslant 80$$
$$H_a: \mu < 80$$

利用一个容量为100的样本，总体标准差为12。取 $\alpha = 0.01$，对下面的每种抽样结果，计算 p-值，并陈述你的结论。

a. $\bar{x} = 78.5$　b. $\bar{x} = 77$　c. $\bar{x} = 75.5$　d. $\bar{x} = 81$

14. 考虑下面的假设检验

$$H_0: \mu = 22$$
$$H_a: \mu \neq 22$$

利用一个容量为75的样本，总体标准差为10。取 $\alpha = 0.01$，对下面的每种抽样结果，计算相应的 p-值，并陈述你的结论。

a. $\bar{x} = 23$　b. $\bar{x} = 25.1$　c. $\bar{x} = 20$

应用

16. 在一项针对"大学毕业生信用卡使用情况"的研究报告中，认为大学毕业生信用卡余额的均值为3 173美元（*Sallie Mae*，2009年4月）。这一数字达到历史新高，与五年前相比增加了44%。假设目前要进行一项研究，确定是否可以得出结论认为毕业生信用卡余额的均值与2009年4月的报告相比是持续增加的。根据已有的研究，令总体标准差为 $\sigma = 1\,000$ 美元。

a. 提出原假设和备择假设。

b. 由180名大学毕业生组成一个样本，信用卡余额的样本均值为3 325美元。求 p-值是多少？

c. 在 $\alpha = 0.05$ 的显著性水平下，你的结论是什么？

18. 1999～2003年，美国多样化股票互助基金的年平均收益为4.1%（《商业周刊》，2004年1月26日）。一名研究者想要通过假设检验来确定，中型成长型基金的年平均收益与美国多样化股票互助基金的年平均收益在同一时期是否存在显著差异。

a. 提出假设，用于确定中型成长型基金年平均收益是否与美国多样化股票基金的年平均收益存在差异。

b. 由40只中型成长型基金组成一个样本，年收益的样本均值 $\bar{x}$ 为3.4%。根据已有的研究，可以假定已知中型成长型基金的总体标准差 $\sigma = 2\%$。利用样本结果，计算假设检验的检验统计量和 p-值。

c. 在 $\alpha = 0.05$ 的显著性水平下，你的结论是什么？

20. 在美国，家庭每个月的互联网账单的均值为32.79美元（CNBC，2006年1月18日）。由南部一个州的50个家庭组成的样本显示，样本均值为 $\bar{x} = 30.63$ 美元。取总体标准差 $\sigma = 5.6$ 美元。

a. 提出假设，用于确定样本数据是否支持作出结论：南部州的家庭每个月互联网账单的均值低于全美32.79美元的平均水平。

b. 检验统计量的值是多少？

c. p-值是多少？

d. 在 $\alpha = 0.01$ 的显著性水平下，你的结论是什么？

22. CNN和ActMedia提供了一个专门的电视频道，向那些在超市收银台前等待结账的顾客播放新闻、广告和短讯。假定在超市收银台前顾客等待结账时间的总体均值为8分钟，并以此为依据决定电视节目的长度。由实际等待时间组成一个样本，并利用样本进行检验，从而判断实际等待时间的均值与此标准之间是否存在差异。

a. 提出这一应用的假设。

b. 由120名购物者组成一个样本，等待时间的均值为8.4分钟。假定总体标准差 $\sigma = 3.2$ 分钟。求 p-值是多少？

c. 在 $\alpha = 0.05$ 的显著性水平下，你的结论是什么？

d. 计算总体均值的95%置信区间。它支持你的结论吗？

9.4 总体均值的检验：σ 未知情形

本节我们说明在 σ 未知情形下如何对总体均值进行假设检验。由于 σ 未知情形与在抽样前无法对总体标准差进行点估计的情形相对应，因此必须利用样本同时估计 σ 和 μ。于是，在 σ 未知情形下对总体均值进行假设检验时，利用样本均值 $\bar{x}$ 估计 μ，用样本标准差 s 估计 σ。

在 σ 未知情形下，假设检验的步骤与第 9.3 节描述的 σ 已知情形下的步骤相同。但是，由于 σ 未知，p-值和检验统计量的计算上稍有不同。在 σ 已知情形下，检验统计量的抽样分布是标准正态分布。然而，在 σ 未知情形下检验统计量的抽样分布是 t 分布。由于根据样本对 μ 和 σ 同时进行估计，t 分布的变异性更强。

我们在第 8.2 节中已经介绍过，在 σ 未知情形下，总体均值的区间估计是基于被称为 t 分布的概率分布。在 σ 未知情形下，总体均值的假设检验仍然是基于 t 分布。对于 σ 未知情形，检验统计量服从自由度为 $n-1$ 的 t 分布。

总体均值假设检验的检验统计量：σ 未知

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{9-2}$$

在第 8 章中我们讲到，t 分布是在假设抽样总体服从正态分布下得到的。然而，研究表明，在样本容量足够大的情形下，可以考虑适当放松这一假设。在本节末，我们将给出实际应用中关于总体分布和样本容量的一些建议。

9.4.1 单侧检验

在 σ 未知情形下，我们考虑关于总体均值单侧检验的例子。一本商务旅行方面的杂志想根据商务旅客总体的评定来划分跨太平洋通道的机场等级。评定标准中最低分为 0，最高分为 10。总体平均等级超过 7 的机场将被认为是提供了优质服务的机场。杂志的职员在每一个机场选取 60 名商务旅客组成一个样本，得到他们的评级数据。在伦敦希斯罗机场的样本中，样本均值 $\bar{x}=7.25$ 分，样本标准差 $s=1.052$ 分。数据能否表明希斯罗机场可以被认为是提供了优质服务的机场？数据存于文件 AirRating 中。

我们想要进行一个假设检验，拒绝 H_0 将会得出结论：希斯罗机场的总体平均等级大于 7 分。因此，要求采用上侧检验，其中 $H_a:\mu>7$。上侧检验的原假设和对立假设如下：

$$H_0:\mu \leqslant 7$$
$$H_a:\mu > 7$$

检验中取显著性水平 $\alpha=0.05$。

利用式（9-2），$\bar{x}=7.25$，$\mu_0=7$，$s=1.052$ 和 $n=60$，检验统计量的值为

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7.25 - 7}{1.052/\sqrt{60}} = 1.84$$

t 分布的自由度为 $n-1=60-1=59$。由于该检验是一个上侧检验，因此 p-值 $=P(t\geqslant 1.84)$，即 t 分布曲线下统计量的值 $t=1.84$ 右侧的面积。

绝大多数教科书里提供的 t 分布表都没有详细到能使我们确定精确的 p-值的地步，比如，与 $t=1.84$ 相对应的 p-值。例如，利用附录 B 中的表 2，自由度为 59 的 t 分布给出了下面的信息。

上侧面积	0.20	0.10	0.05	0.025	0.01	0.005
t 值（自由度为 59）	0.848	1.296	1.671	2.001	2.391	2.662

↑ $t=1.84$

我们看到，$t=1.84$ 介于 1.671 ~ 2.001。虽然，该表不能给出精确的 p-值，但是“上侧面积”这一行中的数值说明：p-值一定大于 0.025 而小于 0.05。在 $\alpha=0.05$ 的显著性水平下，我们根据这些信息足以作出拒绝原假设的决策，并且得到结论：应该将希斯罗机场评定为提供了优质服务的机场。

使用 t 分布表计算 p 值非常烦琐，并且只能给出近似的值。我们将在附录 F 中说明如何用 Excel 或者 Minitab 计算 p 值。在希斯罗机场的假设检验中，$t=1.84$，利用 Excel 或者 Minitab 所计算出的精确的 p-值为 0.035 4。由于 $0.0354<0.05$，我们拒绝原假设并作出结论：认为应该将希斯罗机场评定为提供了优质服务的机场。

在 σ 未知情形下，使用临界值法同样可以得出原假设是否被拒绝的结论。自由度为 59 的 t 分布中，上侧面积 $\alpha=0.05$ 对应于临界值为 $t_{0.05}=1.671$，因此，使用临界值法的拒绝法则为：如果 $t\geqslant 1.671$，则拒绝 H_0。由于 $t=1.84>1.671$，所以拒绝 H_0。希斯罗机场被评定为提供了优质服务的机场。

9.4.2　双侧检验

作为 σ 未知情形下总体均值的双侧假设检验的例子，考虑 Holiday Toys 假设检验中所面临的问题。企业生产产品并通过超过 1 000 多家的零售商分销其产品。在为即将到来的冬季制定生产规模计划时，Holiday Toys 必须在知道零售层面的实际需求量前确定每种产品生产的数量。对本年度最重要的一种新款玩具，Holiday Toys 的市场负责人预计平均每家零售商的需求量为 40 个。在根据这一估计作出最后的生产决策之前，Holiday Toys 决定对 25 个零售商组成的样本进行调查，以便得到有关这种新款玩具需求量的更多信息。在向每个零售商提供有关这种新款玩具的特征、成本以及建议零售价格等方面的信息后，要求每个零售商给出一个预计的订货量。

令 μ 表示零售商订货量的总体均值，根据样本数据进行以下双侧假设检验：

$$H_0:\mu=40$$

$$H_a:\mu\neq 40$$

如果不能拒绝 H_0，那么 Holiday Toys 将继续根据市场负责人的估计制定生产计划，认为每家零售商订货量的总体均值为 40 个。然而，如果拒绝 H_0，那么 Holiday Toys 将会立即重新评估产品的生产计划。因为当零售商订货量的总体均值小于或者大于预计数量时，Holiday Toys 就要重新对其生产计划进行评价，所以采用双侧假设检验。由于没有历史数据可以利用（这是一种新产品），所以必须从样本数据出发用 $\bar{x}$ 和 s 估计总体均值 μ 和总体标准差。

由 25 家零售商组成一个样本，数据存在文件 Orders 中，样本均值 $\bar{x}=37.4$，标准差 $s=11.79$ 个。在利用 t 分布之前，为了了解总体分布的形状，分析人员绘制了样本数据的直方图。样本数据的直方图表明，没有偏斜或者异常点存在，于是分析人员认为采用自由度 $n-1=24$ 的 t 分布是合适的。利用式（9-2），$\bar{x}=37.4$，$\mu_0=40$，$s=11.79$ 和 $n=25$，检验统计量的值为

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{37.4-40}{11.79/\sqrt{25}}=-1.10$$

由于这是一个双侧检验，因此 p-值是 t 分布曲线下 $t\leqslant -1.10$ 部分面积的 2 倍。利用附录 B 中表 B-2，自由度为 24 的 t 分布提供以下信息。

上侧面积	0.20	0.10	0.05	0.025	0.01	0.005
t 值（自由度 24）	0.857	1.318	1.711	2.064	2.492	2.797

$t=1.10$

t 分布表中只包含正值（对应于上侧面积）。然而，由于 t 分布是对称的，从而 $t=1.10$ 右侧曲线下方的面积与 $t=-1.10$ 左侧曲线下方的面积相等。我们看到，$t=1.10$ 介于 0.857 ~ 1.318。在“上侧面积”这一行中看到，$t=1.10$ 右侧面积介于 0.10 ~ 0.20，乘以 2 后，p-值一定介于 0.20 ~ 0.40。在 $\alpha=0.05$ 的显著性水平下，我们知道 p-值大于 α。所以，我们不能拒绝 H_0，没有充分的证据可以得出结论：Holiday Toys 应在即将到来冬季改变生产计划。

在附录 F 中说明了如何用 Excel 或者 Minitab 计算检验的 p-值，计算出的 p-值为 0.282 2。在 $\alpha=0.05$ 的显著性

水平下，由于0.2822>0.05，所以不能拒绝H_0。

也可以将检验统计量与临界值做比较来进行双侧假设检验决策。取$\alpha=0.05$，自由度为24的t分布中双侧检验的临界值为$-t_{0.025}=-2.604$和$t_{0.025}=2.604$。检验统计量的拒绝法则为

$$\text{如果 } t \leqslant -2.604 \text{ 或者 } t \geqslant 2.604\text{，则拒绝 } H_0$$

因为检验统计量$t=-1.10$，所以不能拒绝H_0。这一结果说明：在即将到来的季节里，Holiday Toys应以期望值$\mu=40$为依据，继续其生产计划。

9.4.3 小结与应用建议

表9-3给出了在σ未知的情形下对总体均值进行假设检验的一个总结。与σ已知情形的主要区别在于：计算检验统计量时用s代替σ。因此，检验统计量服从t分布。

表9-3 总体均值假设检验的小结：σ未知

	下侧检验	上侧检验	双侧检验
假设	$H_0:\mu\geqslant\mu_0$ $H_a:\mu<\mu_0$	$H_0:\mu\leqslant\mu_0$ $H_a:\mu>\mu_0$	$H_0:\mu=\mu_0$ $H_a:\mu\neq\mu_0$
检验统计量	$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
拒绝规则：p-值法	如果p-值$\leqslant\alpha$，则拒绝H_0	如果p-值$\leqslant\alpha$，则拒绝H_0	如果p-值$\leqslant\alpha$，则拒绝H_0
拒绝规则：临界值法	如果$t\leqslant -t_\alpha$，则拒绝H_0	如果$t\geqslant t_\alpha$，则拒绝H_0	如果$t\leqslant -t_{\alpha/2}$或者$t\geqslant t_{\alpha/2}$，则拒绝H_0

本节中假设检验方法的应用依赖于抽样总体的分布以及样本容量。如果总体是正态分布，则在任意样本容量下，本节所介绍的假设检验给出的都是精确的结果。如果总体不是正态分布，则这些方法是近似的。尽管如此，我们发现大多数情形下，当样本容量$n\geqslant30$时都能给出满意的结果。如果总体近似服从正态分布，则在小样本容量下（比如：$n<15$）仍可以得到满意的结果。当总体存在高度偏斜或者有异常点时，建议样本容量应在50以上。

练习

方法

☆24. 考虑以下假设检验

$$H_0:\mu=18$$
$$H_a:\mu\neq18$$

一个容量为48的样本中，样本均值为$\bar{x}=17$，样本标准差为$s=4.5$。

a. 计算检验统计量的值。

b. 根据t分布表（附录B中表2）计算p-值的范围。

c. $\alpha=0.05$时，你的结论如何？

d. 临界值法的拒绝法则是什么？你的结论如何？

26. 考虑以下假设检验

$$H_0:\mu=100$$
$$H_a:\mu\neq100$$

根据一个容量为65的样本，取$\alpha=0.05$，对如下每种抽样结果计算相应的p-值，并给出你的结论。

a. $\bar{x}=103$和$s=11.5$

b. $\bar{x}=96.5$和$s=11.0$

c. $\bar{x}=102$和$s=10.5$

应用

28. 一些股东在提出一项否决议案时，主张CEO的平均任期至少为9年。据《华尔街日报》对企业的一项调查发现，CEO任期的样本均值为$\bar{x}=7.27$年，标准差为$s=6.38$年（《华尔街日报》，2007年1月2日）。

a. 提出假设，用于挑战这些股东主张的合理性。

b. 假设样本中有85家企业。假设检验的p-值是多少？

c. 在$\alpha=0.01$的显著性水平下，你的结论是什么？

30. 已婚男性用于照料子女时间的均值为每周6.4个小时（《时代周刊》，2012年3月）。一个家庭护理方面的专业小组想要研究确定本地区已婚男性每周用于照料子女的时间与报道中的均值6.4个小时是否有差异，假定你是该小组的一名成员。由40名

已婚夫妇组成一个样本并搜集丈夫每周用于照料子女时间的数据，样本数据存放在名为 ChildCare 的文件中。

a. 如果你所在的小组想要确定你所属地区已婚男性用于照顾子女的时间的总体均值是否与《时代周刊》所报道的均值有差异，应该提出怎样的假设？

b. 样本均值和 p-值是多少？

c. 选取一个合适的显著性水平，据此你得到怎样的结论？

32. 据全美汽车经销商联合会报告，二手汽车的平均价格为 10 192 美元。由堪萨斯市二手车市场最近售出的 50 辆二手车组成一个样本，该市场的一名管理人员试图确定，堪萨斯市二手车的平均价格与全美的平均水平是否存在差异。50 辆二手车的样本数据存放在名为 UsedCars 的文件中。

a. 提出假设，用于确定该二手车市场的平均价格与全美的平均水平是否存在差异。

b. p-值是多少？

c. 在 $\alpha = 0.05$ 的显著性水平下，你的结论是什么？

34. 琼的 Nursery 公司专门为居民区的环境美化提供设计服务。需要依据种植的树木、灌木等数量来估计一个环境美化项目所需的人工成本。管理者估算认为，种植中型树木需要花费的人工成本为 2 个小时。将上个月 10 次种植树木所花费的实际工作时间组成一个样本，其数据如下（单位：小时）。

1.7　1.5　2.6　2.2　2.4

2.3　2.6　3.0　1.4　2.3

在 0.05 的显著性水平下，检验植树所需时间的均值是否与 2 小时有差异。

a. 提出原假设和备择假设。

b. 计算样本均值。

c. 计算样本标准差。

d. p-值是多少？

e. 你的结论是什么？

9.5 总体比率

在本节中，我们说明如何对总体比率 p 进行假设检验。令 p_0 代表总体比率的假设值，关于总体比率的假设检验有以下三种形式：

$$H_0: p \geq p_0 \qquad H_0: p \leq p_0 \qquad H_0: p = p_0$$

$$H_a: p < p_0 \qquad H_a: p > p_0 \qquad H_a: p \neq p_0$$

称第一种形式为下侧检验，称第二种形式为上侧检验，称第三种形式为双侧检验。

在对总体比率进行假设检验时，是基于样本比率 $\bar{p}$ 与总体比率的假设值 p_0 之差来进行的。假设检验所使用的方法与对总体均值进行假设检验时所使用的方法相似，唯一的不同之处是我们利用样本比率和标准误差来计算检验统计量。然后，利用 p-值法或者临界值法确定是否拒绝原假设。

我们以 Pine Greek 高尔夫球场面临的问题为例。去年，在 Pine Greek 打球的人当中有 20% 是女性。为了增加女性球员的比率，Pine Greek 推出了一项特别的促销活动以吸引更多的女性参加高尔夫运动。在这种特定的促销活动实施一个月以后，高尔夫球场的管理者要求通过统计研究以确定 Pine Greek 的女性高尔夫球员所占比率是否上升。由于研究的目的是确定女性高尔夫球员所占比率是否上升，因此上侧检验比较合适，备择假设为 $H_a: p > 0.20$。Pine Greek 假设检验的原假设和备择假设如下：

$$H_0: p \leq 0.20$$

$$H_a: p > 0.20$$

如果能够拒绝 H_0，则检验结果会对女性高尔夫球员所占比率上升的结论给予统计上的支持，从而说明促销活动是有效的。在进行假设检验时，球场管理者设定显著性水平 $\alpha = 0.05$。

假设检验的下一步是选取一个样本，并计算适当的检验统计量的值。为了说明在 Pine Greek 上侧检验中这一步骤是如何进行的，我们从讨论总体比率假设检验的一般形式出发，说明如何计算检验统计量的值。$\bar{p}$ 是总体参数 p 的点估计量，$\bar{p}$ 的抽样分布是计算检验统计量的基础。

当作为等式的原假设为真时，$\bar{p}$ 的期望值等于假设值 p_0，即 $E(\bar{p}) = p_0$。$\bar{p}$ 的标准误差为

$$\sigma_{\bar{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$$

在第7章中我们曾说过，如果 $np \geq 5$ 并且 $n(1-p) \geq 5$，则 $\bar{p}$ 的抽样分布近似服从正态分布[⊖]。这时，在实践中认为下式

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} \tag{9-3}$$

服从标准正态概率分布，其中 $\sigma_{\bar{p}} = \sqrt{p_0(1-p_0)/n}$。对总体比率进行假设检验时，使用标准正态随机变量 z 为检验统计量。

总体比率假设检验的检验统计量

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{9-4}$$

现在我们计算 Pine Greek 假设检验中的检验统计量。假设选取400名高尔夫球员组成一个随机样本，其中100人是女性。数据文件 WomenGolf 存在本书配套光盘中。样本中女性高尔夫球员的比率为

$$\bar{p} = \frac{100}{400} = 0.25$$

根据式（9-4），检验统计量的值为

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.25 - 0.20}{\sqrt{\frac{0.2(1-0.2)}{400}}} = \frac{0.05}{0.02} = 2.50$$

由于 Pine Greek 假设检验是上侧检验，因此 p-值是 z 大于或者等于 2.50 的概率，即与 $z \geq 2.50$ 相对应区域的面积。利用标准正态概率表，得到 $z = 2.50$ 左侧的面积为 0.993 8。从而，Pine Greek 检验的 p-值为 $1.000\,0 - 0.993\,8 = 0.006\,2$。图9-7给出了 p-值的计算过程。

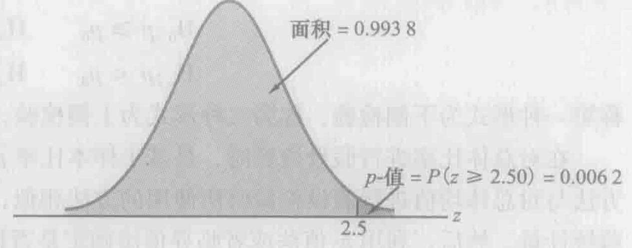


图 9-7 Pine Greek 假设检验中 p-值的计算

球场管理者设定显著性水平 $\alpha = 0.05$。p-值 $= 0.006\,2 < 0.05$，有充分的统计依据在 0.05 的显著性水平下拒绝 H_0。于是，检验给出统计上的依据支持得出结论，认为特定的促销活动能增加 Pine Greek 女高尔夫球运动员的比率。

也可以采用临界值法作出是否拒绝原假设的决策。与正态概率分布上侧面积 0.05 相对应的临界值为 $z_{0.05} = 1.645$。因此，利用临界值法时的拒绝法则为：如果 $z \geq 1.645$ 则拒绝 H_0。由于 $z = 2.50 > 1.645$，因此拒绝 H_0。

我们再次发现，p-值法和临界值法得到的假设检验结论是相同的，但 p-值法提供了更多的信息。p-值 $= 0.006\,2$ 表明，对任何大于或等于 0.006 2 的显著性水平，原假设都能被拒绝。

9.5.1 小结

对总体均值进行假设检验的方法和对总体比率进行假设检验的方法非常相似。虽然我们仅仅说明了如何对总体比率的上侧检验进行假设检验，但是类似的方法也适用于下侧检验和双侧检验。表9-4给出了对总体比率假设检验的一个小结。我们假定 $np \geq 5$ 并且 $n(1-p) \geq 5$，于是 $\bar{p}$ 的抽样分布近似服从正态概率分布。

⊖ 在绝大多数对总体比率进行假设检验的实际应用中，样本容量是足够大的，从而可以利用正态近似。$\bar{p}$ 的精确抽样分布是离散型分布，$\bar{p}$ 取每个值的概率由二项分布给出。因此，小样本情形下，当不能采用正态近似时，假设检验要稍微复杂一些。

表 9-4 总体比率假设检验的小结

	下侧检验	上侧检验	双侧检验
假设	$H_0: p \geqslant p_0$ $H_a: p < p_0$	$H_0: p \leqslant p_0$ $H_a: p > p_0$	$H_0: p = p_0$ $H_a: p \neq p_0$
检验统计量	$z=\dfrac{\bar{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z=\dfrac{\bar{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$z=\dfrac{\bar{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$
拒绝法则：p-值法	如果 p-值$\leqslant\alpha$，则拒绝 H_0	如果 p-值$\leqslant\alpha$，则拒绝 H_0	如果 p-值$\leqslant\alpha$，则拒绝 H_0
拒绝法则：临界值法	如果 $z \leqslant -z_\alpha$，则拒绝 H_0	如果 $z \geqslant z_\alpha$，则拒绝 H_0	如果 $z \leqslant -z_{\alpha/2}$ 或者 $z \geqslant z_{\alpha/2}$，则拒绝 H_0

练习

方法

☆36. 考虑下面的假设检验

$$H_0: p \geqslant 0.75$$
$$H_a: p < 0.75$$

抽取300项组成一个样本，取 $\alpha=0.05$，对下面的每种样本结果，计算 p-值并陈述你的结论。

a. $\bar{p}=0.68$　　b. $\bar{p}=0.72$

c. $\bar{p}=0.70$　　d. $\bar{p}=0.77$

应用

☆38.《消费者报告》的一项研究显示，有64%的超市购物者认为超市自有品牌与国家名牌一样好。某国家名牌番茄酱的制造商为了调查该研究结果对其产品是否适用，抽取超市购物者组成一个样本，并询问样本中的购物者是否认为：超市自有品牌的番茄酱与国家名牌的番茄酱一样好。

a. 提出假设，用于确定超市购物者中认为超市自有品牌子与国家名牌的番茄酱一样好的比率是否与64%有差异。

b. 如果在由100名超市购物者组成的样本中，有52人认为超市自有品牌与国家名牌的番茄酱一样好，那么检验的 p-值是多少？

c. 在 $\alpha=0.05$ 时，你能得出怎样的结论？

d. 国家名牌番茄酱的制造商对这个结论满意吗？请作出解释。

40. 2008年，有46%的企业雇主会向员工赠送节日礼物。在2009年的一项调查中，有35%的雇主打算向员工赠送节日礼物（Radio WEZV，Myrtle Beach，SC，2009年11月11日）。假定由60名雇主组成一个样本用于调查。

a. 在2009年的调查中有多少名雇主打算向员工赠送节日礼物？

b. 假定样本中的雇主确实会像他们最初承诺的那样向员工赠送节日礼物。请用假设检验确定与2008年相比是否雇主向员工赠送节日礼物的比率有所降低，并计算 p-值。

c. 取显著性水平 $\alpha=0.05$，你能够得出雇主赠送节日礼物的比率有所下降的结论吗？要想得到这样的结论，至少应该选取多大的显著性水平？

42. 据内华达大学物流管理中心报道，在美国售出的商品中有6%会遭到退货（《商业周刊》，2007年1月15日）。一月，休斯敦一家百货公司抽取80件售出商品组成样本，发现其中12件遭到退货。

a. 在休斯敦百货公司售出商品中，求遭到退货的总体比率的点估计。

b. 建立休斯敦百货公司售出商品的退货比率的95%置信区间。

c. 休斯敦百货公司的退货比率与全国的退货比率存在显著差异吗？为你的答案提供统计上的支持。

44.《商业周刊》在封面故事中公布了关于美国人睡眠习惯的信息（《商业周刊》，2004年1月26日）。作者强调，睡眠不足导致了包括高速路车祸死亡在内的许多问题。成人驾驶员中有51%的人承认在驾驶中感觉困倦。某研究者假设这一问题在夜班驾驶员中尤为严重。

a. 提出假设，用于确定是否在由夜班驾驶员组成的总体中有超过51%的人在驾驶时感到困倦。

b. 由400名夜班驾驶员组成一个样本，数据存放在名为 Drowsy 的文件中。请识别出那些承认自己在驾驶中感到困倦的人，并计算样本比率和 p-值。

c. 在 $\alpha=0.01$ 的显著性水平下，你的结论是什么？

9.6 假设检验与决策

在本章的以前各节中，我们介绍的假设检验在应用中被称之为显著性检验。在提出原假设和备择假设之后，我们选取样本，计算检验统计量的值和相应的 p 值。我们将 p 值与控制第一类错误概率的 α 相比，其中 α 被称作检验的显著性水平。如果 p 值 $\leqslant\alpha$，则我们作出结论“拒绝 H_0”，并称结果是显著的。否则，我们将作出“不能拒绝 H_0”的结论。在显著性检验中，我们控制了发生第一类错误的概率，但没有控制发生第二类错误的概率。因此，我们给出的结论是“不能拒绝 H_0”，而不是“接受 H_0”。因为，后者会使我们在 H_0 为假时接收 H_0，从而承担发生第二类错误的风险。“不能拒绝 H_0”的结论，从统计证据上讲是非结论性的，这通常意味着将决策或行动推迟，待经过更深入的研究和检验之后再作决策。

然而，如果假设检验的目的是：当 H_0 为真时，作出一种决策；当 H_a 为真时，作出另一种决策。那么，不论是作出“拒绝 H_0”还是“不能拒绝 H_0”的结论，决策者都想要或者在某些情形下不得不作出决策。这时，统计学家往往建议控制发生第二类错误的概率。当同时控制发生第一类错误和第二类错误的可能性时，假设检验的结论要么是“接受 H_0”或者“拒绝 H_0”。对第一种情形，结论认为 H_0 为真；而在第二种情形，结论认为 H_a 为真。在任何一种情况下，都必须作出决策并采取相应的措施。

货物接收的抽样问题，就是假设检验在决策中一个很好的应用。第 20 章中将对这一问题展开更加深入的讨论。例如，对供应商的一批电池，某名质量控制管理人员必须决定是接受这批货物还是因其质量差而将货物退还供应商。假定，设计规格中要求供应商电池的平均使用寿命至少为 120 个小时。为了评估这批货物的质量，我们选取 36 节电池组成样本进行检验，根据样本来决定是接受这批货物还是因其质量差而将这批货物退还供应商。令 μ 代表这批电池使用寿命的均值（单位：小时），对总体均值建立如下形式的原假设和备择假设。

$$H_0: \mu \geqslant 120$$

$$H_a: \mu < 120$$

如果拒绝 H_0，则作出备择假设为真的结论。这一结论表明，将这批货物退还给供应商是明智之举。但是，如果不能拒绝 H_0，决策者仍需确定采取某种措施。所以，当没有直接得出“H_0 为真”的结论，而仅仅是“不能拒绝 H_0”时，决策者将认为这批货物的质量符合标准而决定接收这批货物。

在这种决策中，我们建议将假设检验过程扩展，对发生第二类错误的概率予以控制。由于当不能拒绝 H_0 时，必须作出决策并采取措施，所以有关发生第二类错误的概率的信息对我们很有帮助。在第 9.7 节和第 9.8 节中，我们将说明如何计算发生第二类错误的概率，如何调整样本容量帮助控制发生第二类错误的概率。

9.7 计算第二类错误的概率

在本节中，我们说明在总体均值的假设检验中如何计算发生第二类错误的概率。以第 9.6 节中所描述的货物接收问题为例，我们对该过程进行演示。电池使用寿命（单位：小时）均值的原假设和备择假设分别为 $H_0: \mu \geqslant 120$ 和 $H_a: \mu < 120$。如果拒绝 H_0，则因电池使用寿命的均值小于规格所要求的 120 个小时，而决定将这批货物退回供应商。如果不能拒绝 H_0，那么就决定接收这批货物。

假定假设检验时所采用的显著性水平 $\alpha = 0.05$，在 σ 已知的情形下，检验统计量为

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{x} - 120}{\sigma / \sqrt{n}}$$

根据临界值方法以及 $z_{0.05} = 1.645$，下侧检验的拒绝规则为

$$\text{如果 } z \leqslant -1.645\text{，则拒绝 } H_0$$

假定选取 36 节电池组成一个样本，根据已有的检验可以假定总体标准差已知，为 $\sigma = 12$ 个小时。拒绝规则表明，当

$$z = \frac{\bar{x} - 120}{12 / \sqrt{36}} \leqslant -1.645$$

拒绝 H_0。上式中关于 $\bar{x}$ 求解可知，当

$$\bar{x} \leqslant 120 - 1.645\left(\frac{12}{\sqrt{36}}\right) = 116.71$$

拒绝 H_0。这意味着，当 $\bar{x} > 116.71$ 时，我们将作出接收这批货物的决策。利用这些信息，我们可以计算发生第二类错误相的概率。首先，当货物均值的真值小于 120 小时而我们却作出接受 $H_0: \mu \geqslant 120$ 的决定时，我们就犯了第二类错误。因而，为了计算发生第二类错误的概率，我们必须选择一个小于 120 小时的 μ 值。例如，假定电池寿命的均值 $\mu = 112$ 小时，该批货物的质量差。那么，在 $\mu = 112$ 确实是真但我们却接受了 $H_0: \mu \geqslant 120$，因此发生第二类错误的概率是多少呢？它应该是当 $\mu = 112$ 时样本均值 $\bar{x}$ 大于 116.71 的概率。

图 9-8 给出了当均值 $\mu = 112$ 时 $\bar{x}$ 的抽样分布，其上侧阴影部分的面积为 $\bar{x} > 116.71$ 的概率。根据标准正态分布，当 $\bar{x} = 116.71$ 时

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{116.71 - 112}{12 / \sqrt{36}} = 2.36$$

由标准正态概率表可知，当 $z = 2.36$ 时其上侧面积是 $1.0000 - 0.9909 = 0.0091$，0.0091 是当 $\mu = 112$ 时发生第二类错误的概率。令 β 表示发生第二类错误的概率，可见当 $\mu = 112$ 时，$\beta = 0.0091$。从而，我们得出结论：如果总体均值为 112 小时，则发生第二类错误的概率仅为 0.0091。

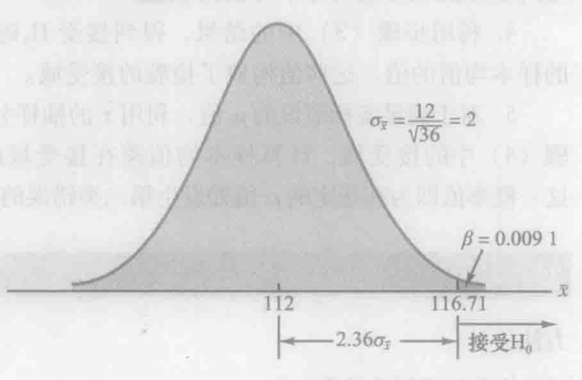


图 9-8 当 $\mu = 112$ 时发生第二类错误的概率

对其他小于 120 的 μ 值，我们可以重复该计算过程，求出每一个 μ 值下发生第二类错误的概率。例如，假定货物中电池使用寿命的均值 $\mu = 115$ 小时。由于当 $\bar{x} > 116.71$ 时，我们接受 H_0，所以 $\mu = 115$ 时 z 值为

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{116.71 - 115}{12 / \sqrt{36}} = 0.86$$

查标准正态概率表可得，当 $z = 0.86$ 时，标准正态分布上侧面积为 $1.0000 - 0.8051 = 0.1949$。因此，当均值的真值 $\mu = 115$ 时，发生第二类错误的概率 $\beta = 0.1949$。

表 9-5 列出了当 μ 取某些小于 120 的值时发生第二类错误的概率。我们注意到：随着 μ 逐渐增加到 120，发生第二类错误的概率也随之增大，达到上限 0.95。相反，随着 μ 值比 120 越来越小，发生第二类错误的概率也逐渐减小。不出所料，当总体均值的真值在原假设的值 $\mu = 120$ 附近时，发生第二类错误的概率很大，但是当总体均值的真值在远离原假设的值 $\mu = 120$ 的下方时，则发生第二类错误的概率很小。[⊖]

表 9-5 货物接收假设检验问题中发生第二类错误的概率

μ 值	$z = \frac{116.71 - \mu}{12/\sqrt{36}}$	第二类错误的概率 β	功效 $(1-\beta)$
112	2.36	0.0091	0.9909
114	1.36	0.0869	0.9131
115	0.86	0.1949	0.8051
116.71	0.00	0.5000	0.5000
117	−0.15	0.5596	0.4404
118	−0.65	0.7422	0.2578
119.999	−1.645	0.9500	0.0500

当 H_0 为假时，作出拒绝 H_0 的正确结论的概率称作检验的**功效**（power）。对于给定的 μ 值，功效为 $1 - \beta$，即作出拒绝原假设正确结论的概率等于 1 减去发生第二类错误的概率。功效的值列示在表 9-5 中，根据这些值绘制与每一个 μ 值相对应的功效，所得曲线称为**功效曲线**（power curve），如图 9-9 所示。注意，功效曲线渐近于原假

⊖ 正如表 9-5 所示，第二类错误的概率依赖于总体均值 μ 的值。对 μ_0 附近的 μ，发生第二类错误的概率很高。

设为假时的μ值。对任意一个μ值，功效曲线的高度代表了当H_0为假时作出拒绝H_0正确结论的概率。[⊖]

总之，对总体均值进行假设检验时，可以按照如下所给的流程逐步计算得到发生第二类错误的概率。

1. 确立原假设和备择假设。

2. 在显著性水平α下，根据临界值方法确定临界值并建立检验的拒绝规则。

3. 利用（2）中所得的拒绝规则，求解与检验统计量的临界值相对应的样本均值的取值。

4. 利用步骤（3）中的结果，得到接受H_0时所对应的样本均值的值，这些值构成了检验的接受域。

5. 对于满足备择假设的μ值，利用$\bar{x}$的抽样分布和步骤（4）中的接受域，计算样本均值落在接受域的概率。这一概率值即为在选定的μ值处发生第二类错误的概率。

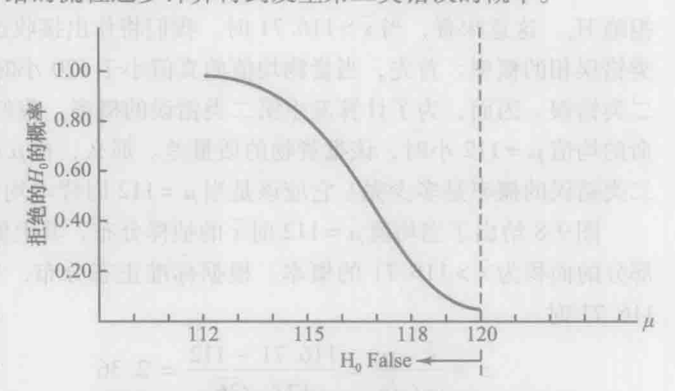


图9-9　货物接收假设检验问题中的功效曲线

练　习

方法

☆46. 考虑如下假设检验。

$$H_0: \mu \geqslant 10$$
$$H_a: \mu < 10$$

样本容量为120，假定已知总体标准差$\sigma=5$，取$\alpha=0.05$。

a. 如果总体均值为9，则根据样本均值我们作出不能拒绝H_0的结论的概率为多大？

b. 如果实际的总体均值为9，而我们却得出H_0：$\mu \geqslant 10$为真的结论，则这时会发生哪一类错误？

c. 如果实际的总体均值为8，则发生第二类错误的概率为多少？

应用

48. Fowle市场研究有限公司假定电话调查可在15分钟以内结束，并据此向顾客收费。如果调查所需时间超过该值，则需要加收额外费用。由35个调查电话组成一个样本，总体标准差为4分钟，利用样本均值检验原假设$H_0: \mu \leqslant 15$。取显著性水平为0.01。

a. 你如何解释该问题中的第二类错误？它对公司有什么影响？

b. 当实际所用时间的均值$\mu=17$分钟时，发生第二类错误的概率是多少？

c. 当实际所用时间的均值$\mu=18$分钟时，发生第二类错误的概率是多少？

d. 画出该检验中功效曲线的一般形状。

50. *Young Adult*杂志对其订阅者的年龄均值作出如下假设

$$H_0: \mu = 28$$
$$H_a: \mu \neq 28$$

a. 这种情况下，发生第二类错误意味着什么？

b. 假设已知总体标准差$\sigma=6$岁，样本容量为100，取$\alpha=0.05$，则当μ值分别等于26，27，29和30时，接受H_0的概率为多大？

c. 当$\mu=26$时，功效为多少？这说明什么？

52. 参见习题48。假定公司选取50次调查组成一个样本，重复（b）和（c）。随着样本容量的增加，你发现发生第二类错误的概率将如何变化？

9.8　对总体均值进行假设检验时样本容量的确定

假定要对总体均值进行假设检验，检验中使用者事先指定显著性水平，以确定检验中发生第一类错误的概率。通过控制样本容量，使用者也可以控制发生第二类错误的概率。下面，我们说明在总体均值的下侧检验中如何确定样本容量。

$$H_0: \mu \geqslant \mu_0$$
$$H_a: \mu < \mu_0$$

⊖ 有时，我们利用另一种称之为工作特征曲线的图形来给出发生第二类错误的概率的信息。工作特征曲线给出了接受H_0的概率以及当原假设为假时μ值所对应的β值。由该图可以直接得出发生第二类错误的概率。

图9-10的上半部分是当H_0为真并且$\mu=\mu_0$时$\bar{x}$的抽样分布。对于下侧检验，检验统计量的临界值记做$-z_\alpha$。图上半部分中垂线c与$\bar{x}$相对应。注意，如果当$\bar{x}\leqslant c$时拒绝H_0，则发生第一类错误的概率为α。以z_α表示标准正态分布上侧面积为α时所对应的z值，可利用如下公式计算c

$$c=\mu_0-z_\alpha\frac{\sigma}{\sqrt{n}} \tag{9-5}$$

图9-10的下半部分为当备择假设H_a为真并且$\mu=\mu_a<\mu_0$时$\bar{x}$的抽样分布。阴影区域的面积β恰好是当$\bar{x}>c$时却接受了原假设时决策者发生第二类错误的概率。以z_β表示标准正态分布上侧面积为β时所对应的z值，可利用如下公式计算c。

$$c=\mu_a+z_\beta\frac{\sigma}{\sqrt{n}} \tag{9-6}$$

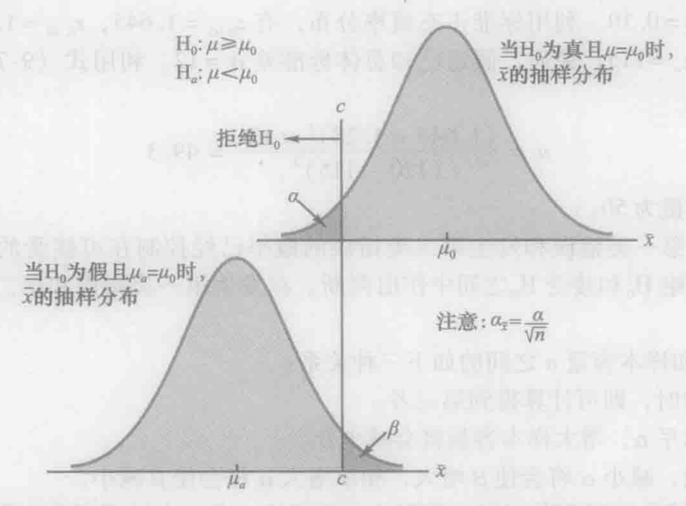


图9-10　给定第一类错误水平（α）和第二类错误水平（β），样本容量的确定

由于我们想要选取一个c值使得当我们拒绝H_0接受H_a时，发生第一类错误的概率等于选取的值α，发生第二类错误的概率等于选取的值β。因此，式（9-5）和（9-6）所给出的c值是相等的。于是，如下表达式一定成立。

$$\mu_0-z_\alpha\frac{\sigma}{\sqrt{n}}=\mu_a+z_\beta\frac{\sigma}{\sqrt{n}}$$

为了确定所需要的样本容量，我们首先用如下方式求解$\sqrt{n}$。

$$\mu_0-\mu_a=z_\alpha\frac{\sigma}{\sqrt{n}}+z_\beta\frac{\sigma}{\sqrt{n}}$$

$$\mu_0-\mu_a=\frac{(z_\alpha+z_\beta)\sigma}{\sqrt{n}}$$

和

$$\sqrt{n}=\frac{(z_\alpha+z_\beta)\sigma}{(\mu_0-\mu_a)}$$

等式两边平方，即得到如下公式用于确定总体均值单侧假设检验的样本容量。

总体均值单侧假设检验中的样本容量

$$n=\frac{(z_\alpha+z_\beta)^2\sigma^2}{(\mu_0-\mu_a)^2} \tag{9-7}$$

式中，z_α代表标准正态分布的上侧面积为α时对应的z值；z_β代表标准正态分布的上侧面积为β时对应的z值；σ为总体的标准差；μ_0为原假设中样本均值的值；μ_a为第二类错误中所采用的总体均值的值。

注意：双侧假设检验中，在式（9-7）中以$z_{\alpha/2}$代替z_α。

虽然式（9-7）的逻辑是用于如图9-10所示的假设检验的，但是对其他形式总体均值的单侧检验也成立。在总体均值的双侧假设检验中，只需在式（9-7）中以$z_{\alpha/2}$代替z_α即可。

我们仍采用第9.6节和第9.7节的货物接收例子。设计规格要求电池使用寿命的均值至少为120个小时。如果$H_0:\mu \geq 120$被拒绝，则该批货物被拒收。我们假定质量控制管理人员要求第一类错误和第二类错误的可接受概率达到如下要求。

关于第一类错误的要求：如果货物中电池寿命的均值为$\mu = 120$，那么甘愿冒$\alpha = 0.05$的风险概率拒绝这批货物。

关于第二类错误的要求：如果货物中电池寿命的均值比规格要求少5小时（即$\mu = 115$），那么甘愿冒$\beta = 0.10$的风险概率接受这批货物。

如上要求是基于管理者的判断得到的。不同的人可能会对此概率作出不同的限制。但是，在确定样本容量之前，必须明确允许发生两类错误概率的大小。

在例子中，$\alpha = 0.05$，$\beta = 0.10$。利用标准正态概率分布，有$z_{0.05} = 1.645$，$z_{0.10} = 1.28$。从关于错误概率的要求中，我们已知$\mu_0 = 120$，$\mu_a = 115$。最后，假定已知总体标准差$\sigma = 12$。利用式（9-7），在货物接收的例子中，我们建议使用的样本容量为

$$n = \frac{(1.645 + 1.28)^2 \times 12^2}{(120 - 115)^2} = 49.3$$

舍入后，建议选取的样本容量为50。

当$n = 50$时，由于发生第一类错误和发生第二类错误的概率已经控制在可接受的水平内，所以在假设检验中，质量控制管理者要在拒绝H_0和接受H_0之间中作出判断。在发生第一类错误和第二类错误的概率在已知允许范围内时，进行相关推断。

我们可以观察到α、β和样本容量n之间的如下三种关系：

1. 当三者中有二者已知时，即可计算得到第三者。
2. 对于给定的显著性水平α，增大样本容量将会减少β。
3. 对于给定的样本容量，减小α将会使β增大，相反增大α将会使β减小。

当未对第二类错误的概率加以控制的时候，我们应该牢记第三条，它说明不能毫无必要地选择太小的显著性水平α。对于给定的样本容量，选择较小的显著性水平意味着将使发生第二类错误的风险增大。缺乏经验的使用者通常认为假设检验中，α的取值越小越好。当我们只关心第一类错误时，确实如此。但不利的是，较小的α值将增大发生第二类错误的概率。

练习

方法

☆54. 考虑如下假设检验。

$$H_0:\mu \geq 10$$
$$H_a:\mu < 10$$

样本容量为120，总体标准差为5，取$\alpha = 0.05$。如果实际上总体均值为9，发生第二类错误的概率为0.291 2。假定当总体的实际均值为9时，研究者想要将发生第二类错误的概率减少到0.10，则应该选取多大的样本容量？

应用

56. 在Hilltop咖啡研究中（见第9.3节），假定项目负责人要求当实际重量少了1盎司（即$\mu_a = 2.937\,5$磅）时Hilltop却不会遭到投诉的概率为0.10，则应采用多大的样本容量？

58. *Young Adult* 杂志对其订阅者的年龄的均值建立如下假设。

$$H_0:\mu = 28$$
$$H_a:\mu \neq 28$$

如果检验中管理者允许发生第二类错误的概率为0.15，而真实年龄的均值为29岁，则应该选取多大的样本容量？假定$\sigma = 6$，取显著性水平为0.05。

小结

假设检验是一种统计方法，它利用样本数据来确定是否拒绝关于总体参数的说法。假设是关于总体参数的两种不同的说法，其中一个称作原假设（H_0），另一个称作备择假设（H_a）。在第9.1节，针对实际应

用中最常发生的情况给出了建立假设的规则。

当根据历史数据或者其他信息可以假定总体标准差已知时，总体均值的假设检验过程以正态分布为依据。当σ未知时，用样本标准差s估计σ，假设检验过程以t分布为依据。在这两种情形下，假设检验结果的质量依赖于总体分布的形式以及样本容量。如果总体服从正态分布，则即使在小样本的情形下，这两种假设检验方法也是适用的。但是，如果总体不服从正态分布，则需要较大的样本容量。第9.3节和第9.4节给出了有关样本容量的一般建议。在对总体比率进行假设检验时，假设检验中利用的检验统计量以标准正态分布为依据。

在所有上述情形下，都可以利用检验统计量的值来计算检验的p-值。p-值用于确定原假设是否被拒绝的概率。当p-值小于或者等于显著性水平α时，我们拒绝原假设。

也可以将检验统计量的值与临界值相比较后得出假设检验的结论。对于下侧检验，如果检验统计量的值小于或者等于临界值，则拒绝原假设。对于上侧检验，如果检验统计量的值大于或者等于临界值，则拒绝原假设。双侧检验包括两个临界值，一个位于抽样分布的上侧，一个位于抽样分布的下侧。在这种情形下，当检验统计量的值小于或等于下侧临界值或者检验统计量的值大于或等于上侧临界值时，则拒绝原假设。

作为假设检验方法的扩展，本章给出了关于第二类错误的分析。在第9.7节，我们介绍了如何计算发生第二类错误的概率。在第9.8节，描述了当同时控制发生第一类错误和第二类错误的概率时如何确定样本容量。

关键术语

null hypothesis 原假设 在假设检验的程序中，尝试性地假定为真的假设。

alternative hypothesis 备择假设 如果原假设被拒绝时，被认为是真的假设。

type II error 第二类错误 当H_0为假时却接受了H_0时所犯的错误。

type I error 第一类错误 当H_0为真时却拒绝了H_0时所犯的错误。

level of significance 显著性水平 当原假设以等式形式为真成立时，发生第一类错误的概率。

one-tailed test 单侧检验 假设检验的一种，当检验统计量的值落在抽样分布的某一侧时，拒绝原假设。

test statistic 检验统计量 一种统计量，它的值用于确定是否拒绝原假设。

p-value p-值 p-值是一个概率值，当根据样本作出拒绝原假设的结论时，p-值是对证据充分程度的一种度量。p-值越小，则越有理由拒绝原假设H_0。对于下侧检验，p-值是检验统计量小于或者等于样本所给出的检验统计量的值的概率。对于上侧检验，p-值是检验统计量大于或者等于样本所给出的检验统计量的值的概率。对于双侧检验，p-值是检验统计量大于或等于检验统计量的绝对值的概率。

critical value 临界值 与检验统计量相比，用于确定是否拒绝H_0的值。

two-tailed test 双侧检验 假设检验的一种，当检验统计量的值落在抽样分布两侧的任一侧时，拒绝原假设。

power 功效 当H_0为假时作出拒绝H_0这一正确结论的概率。

power curve 功效曲线 对所有不满足原假设的总体参数值，功效曲线绘制出了拒绝H_0的概率。功效曲线给出了正确地拒绝原假设的概率。

重要公式

总体均值假设检验的检验统计量：σ已知

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \tag{9-1}$$

总体均值假设检验的检验统计量：σ未知

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}} \tag{9-2}$$

总体比率假设检验的检验统计量

$$z=\frac{\bar{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \tag{9-4}$$

总体均值单侧假设检验的的样本容量

$$n=\frac{(z_\alpha+z_\beta)^2\sigma^2}{(\mu_0-\mu_a)^2} \tag{9-7}$$

用$z_{\alpha/2}$代替z_α，即为双侧检验的公式。

补充练习

60. 某种产品的生产线上，产品充入重量的均值为16盎司。过多或过少都会导致严重的后果，并且一经发现就要求操作者立即关闭生产线对填充机进行重新调试。根据历史数据，假设总体标准差 $\sigma=0.8$ 盎司。质量控制监督员每小时检验30件产品，并同时确定该生产线是否需要停产进行调整。取显著性水平 $\alpha=0.05$。
   a. 提出用于质量控制的假设检验。
   b. 如果样本均值 $\bar{x}=16.32$ 盎司，则 p-值是多少？你建议采取什么措施？
   c. 如果样本均值 $\bar{x}=15.82$ 盎司，则 p-值是多少？你建议采取什么措施？
   d. 利用临界值法，求上述假设检验过程中的拒绝规则。重复（a）和（b）中的问题。你得到了相同的结论吗？

62. *Playbill* 是一家全国发行的杂志，它面向那些专心于音乐片和其他戏剧作品的发烧友。在 *Playbill* 的读者总体中，家庭收入的年均值是119 155美元（*Playbill*，2006年1月）。假设总体标准差 $\sigma=20\,700$ 美元。旧金山市民小组断言，海湾地区剧场戏迷观众的均值会更高。由海湾区60名剧场观众组成的一个样本显示，家庭收入的样本均值为126 100美元。
   a. 提出检验的假设，用来确定样本数据是否支持作出结论认为海湾地区戏剧观众的家庭收入的均值比全体 *Playbill* 读者的均值要高。
   b. 基于海湾地区60名剧场观众组成的一个样本，求 p-值是多少？
   c. 取显著性水平 $\alpha=0.01$，你的结论是什么？

64. 美国国家健康统计中心发布的数据显示，2006年女性生育第一个孩子的年龄平均为25.0岁（《华尔街日报》，2009年2月4日）。记者 Sue Shellenbarger 强调，这是近年来女性生育第一个孩子的年龄首次出现下降。最近，选取42名妇女组成一个样本，她们生育第一个孩子的年龄数据存放在名为 FirstBirth 的网络文件中。数据是否表明2006年女性生育第一个孩子的年龄的均值发生了变化？取 $\alpha=0.05$。

66. 据 Florida Gulf Coast 社区的商会广告称，该地点住宅房产的平均成本不超过125 000美元。假设由32处房产所组成的一个样本表明，其样本均值为130 000美元，样本标准差为12 500美元。在0.05的显著性水平下，检验广告的有效性。

68. 2009年12月25日，在西北航空公司飞往密歇根州底特律的航班上，制服了一名试图炸毁飞机的乘客。这名乘客偷偷把爆炸物藏在内衣中混过了机场案件装置中的金属探测仪。美国运输安全局（TSA）建议在美国的大型机场安装全身扫描装置代替金属探测仪。这一提议遭到隐私倡导者的强烈反对，他们认为全身扫描装置侵犯了个人隐私。在2010年1月的5~6号，《今日美国》对542名成人进行了一项民意调查，了解他们当中有多大比率赞同使用全身扫描装置（《今日美国》，2010年1月11日）。民意调查结果显示，有455人认为全身扫描能够提高航空安全，有432人同意使用该装置。
   a. 利用假设检验的方法，确定民意调查数据是否支持作出结论：超过80%的乘客认为全身扫描能够提高航空安全。取 $\alpha=0.05$。
   b. 假定若有75%的乘客赞同使用，则 TSA 会安装并强制使用全身扫描装置。要求你根据民意调查结果进行统计检验，确定 TSA 是否应该安装并强制使用全身扫描装置。由于这是一个非常敏感的决策，因此检验中取显著性水平 $\alpha=0.01$。你有何建议？

70. 虚拟呼叫中心的职员都是一些外地人。大部分地区代理每小时收入为10~15美元（含保险费）或者每小时收入为7~9美元（不含保险费）（《商业周刊》，2006年1月23日）。某地方航空考虑聘用本地代理，但是只有当顾客的满意水平超过80%时才能留用。对本地代理的服务进行检验，在300名顾客组成的样本中有252人满意他们的服务。
   a. 提出检验的假设，用来确定样本数据是否支持作出顾客认为本地代理的服务达到了地方航空的标准的结论。
   b. 顾客满意的百分比的点估计是多少？
   c. 根据样本数据，求 p-值是多少？
   d. 假设检验的结论是什么？取显著性水平 $\alpha=0.05$。

72. Myrtle 海滩的电台称，在纪念日的周末，至少有

90%的旅馆和汽车旅店会客满。如果打算在周末去该风景点的话，电台建议听众提前预订。周六晚，一个由58家旅馆和汽车旅店组成的样本表明，其中有49家贴出了无空房的招牌，有9家有空房间。在看了这一抽样结果后，你对广播电台的说法做何反应？统计检验中，取$\alpha=0.05$，求p-值为多少？

74. 在建筑计划的投标中，Shorney建筑公司假定每天每名建筑工人空闲时间的均值不超过72分钟。根据一个由30名建筑工人组成的样本对此假设进行检验。假设总体的标准差为20分钟。
    a. 提出检验的假设。
    b. 当空闲时间的总体均值为80分钟时，发生第二类错误的概率为多少？
    c. 当空闲时间的总体均值为75分钟时，发生第二类错误的概率为多少？
    d. 当空闲时间的总体均值为70分钟时，发生第二类错误的概率为多少？
    e. 绘出该问题的功效曲线。

76. 用$H_0:\mu=120$和$H_a:\mu\neq120$来检验某一浴用肥皂生产过程是否达到了每批产量为120块的标准。检验中取显著性水平为0.05，令标准差的计划值为5。
    a. 如果当产量的均值下降到每批117块时，企业将以98%的概率作出结论认为产量未达到标准，那么应采用多大的样本容量？
    b. 在（a）中所得的样本容量下，当实际产量的均值分别为每批117，118，119，121，122和123块时，能以多大的概率得出该生产过程正常运行的结论？即在每种情形下，发生第二类错误的概率为多少？

案例9-1 Quality Associates有限公司

Quality Associates是一家为客户提供抽样和统计方法方面建议的咨询公司，这些建议可以用来监控客户的制造流程。在一个应用项目中，一名客户向Quality Associates提供了一个样本，该样本由制造流程正常运行时的800个观察值组成，样本标准差为0.21。假设总体标准差为0.21。Quality Associates建议：持续不断地定期抽取容量为30的随机样本监测制造流程。通过对这些新样本的分析，客户可以迅速知道制造流程的运行状况是否令人满意。当制造流程的运行不能令人满意时，可以采取纠正措施予以解决。设计规格要求制造流程的均值为12，Quality Associates建议采用如下形式的假设检验。

$$H_0:\mu = 12$$

$$H_a:\mu \neq 12$$

只要H_0被拒绝，就应采取纠正措施。

下表为第一天对制造流程采用这种新的的统计控制方法时，每隔一小时收集的样本数据。这些数据存放在名为Quality的数据集中。

管理报告

1. 对每个样本在0.01的显著性水平下进行假设检验，并且确定，如果需要的话，应该采取怎样的措施？给出每一检验的检验统计量和p-值。

2. 计算四个样本中每个样本的标准差。假设总体标准差为0.21，这样做是否合理？

3. 计算样本均值$\bar{x}$在$\mu=12$附近的一个范围，使得只要样本均值在这个范围内，则认为制造流程的运行状况是令人满意的。如果$\bar{x}$超过上限或低于下限，则需采取纠正措施。在质量控制目标中，这类界限被称为上侧或下侧控制限。

4. 当显著性水平增大时，这意味着什么？如果增大显著性水平，哪种错误或误差会增加？

样本1	样本2	样本3	样本4	样本1	样本2	样本3	样本4
11.55	11.62	11.91	12.02	11.90	11.97	12.14	12.11
11.62	11.69	11.36	12.02	11.64	11.71	11.72	12.07
11.52	11.59	11.75	12.05	11.80	11.87	11.61	12.05
11.75	11.82	11.95	12.18	12.03	12.10	11.85	11.64

（续）

样本1	样本2	样本3	样本4	样本1	样本2	样本3	样本4
11.94	12.01	12.16	12.39	11.77	11.84	12.12	11.79
11.92	11.99	11.91	11.65	12.00	12.07	11.60	12.30
12.13	12.20	12.12	12.11	12.04	12.11	11.95	12.27
12.09	12.16	11.61	11.90	11.98	12.05	11.96	12.29
11.93	12.00	12.21	12.22	12.30	12.37	12.22	12.47
12.21	12.28	11.56	11.88	12.18	12.25	11.75	12.03
12.32	12.39	11.95	12.03	11.97	12.04	11.96	12.17
11.93	12.00	12.01	12.35	12.17	12.24	11.95	11.94
11.85	11.92	12.06	12.09	11.85	11.92	11.89	11.97
11.76	11.83	11.76	11.77	12.30	12.37	11.88	12.23
12.16	12.23	11.82	12.20	12.15	12.22	11.93	12.25

案例 9-2 Bayview 大学商科学生的道德行为

在全球经济衰退的2008年和2009年，有许多针对华尔街管理人员、财务经理和其他公司经理人职业道德的控告。当时，有一篇文章认为不道德商业行为的部分根源在于商科学生中盛行欺诈行为（*Chronicle of Higher Education*，2009年2月10日）。文中称，有56%的商科学生承认他们在学期间曾有过欺诈行为；而在非商科学生中，这一比率为47%。

近年来，欺诈行为一直是Bayview大学商学院教务主任关注的一个问题。一些大学教学人员认为，Bayview大学的欺诈行为比其他大学更为普遍；另一些大学教学人员认为，欺诈行为并不是学院的主要问题。为回复这些说法，教务主任委托对目前Bayview大学商科学生欺诈行为进行评估研究。作为研究的一部分，由90名本年即将毕业的商科学生组成一个样本，向他们分发匿名的调查问卷并回答以下问题，得到与三类欺骗行为有关的数据。

在Bayview大学就学期间，你曾经将互联网上的内容拷贝作为自己的工作吗？

是的________ 不是________________

在Bayview大学就学期间，你曾经在考试中抄袭他人的答案吗？

是的________ 不是________________

在Bayview大学就学期间，你曾经将与其他学生合作的项目当作自己独立完成的吗？

是的________ 不是________________

如果对这些问题的回答中有一项或者一项以上为“是的”，则认为学生具有欺骗行为。下面是搜集到的部分数据。整个数据集在本书光盘名为Bayview的文件中。

学生	从互联网抄袭	考试中抄袭	将合作项目当作个人独立完成	性别
1	No	No	No	女
2	No	No	No	男
3	Yes	No	Yes	男
4	Yes	Yes	No	男
5	No	No	Yes	男
6	Yes	No	No	女
⋮	⋮	⋮	⋮	⋮
88	No	No	No	男
89	No	Yes	Yes	男
90	No	No	No	女

管理报告

请向大学教务处主任提交一份报告，总结评估 Bayview 大学商科学生的欺骗行为。在报告中，请确保包括如下内容。

1. 利用描述统计量对数据进行汇总并评论您的结果。

2. 在全体学生、男学生和女学生当中，分别求发生某种类型欺骗行为的学生所占比率的95%置信区间。

3. 进行假设检验，确定是否 Bayview 大学商科学生中发生欺骗行为的比率低于 *Chronicle of Higher Education* 报道的其他大学商科学生？

4. 进行假设检验，确定是否 Bayview 大学商科学生中发生欺骗行为的比率低于 *Chronicle of Higher Education* 报道的其他大学非商科学生？

5. 根据你对数据的分析，你对系主任有哪些建议？

第 10 章

两总体均值和比例的推断

CHAPTER 10

实践中的统计

美国食品与药物管理局

华盛顿特区

美国食品与药物管理局（U. S. Food and Drug Administration，FDA）的职责是通过其药物评估和研究中心（CDER）来保证药物的安全性和有效性。但 CDER 自己并不会对任何新药进行实际检验。试图推出新药的公司有责任对该药进行检验并提交有关该药安全性和有效性的证明，然后 CDER 的统计学家和科学家会对提交的证明进行审查。

为了使一种新药获得批准，公司进行了大量的统计研究来支撑他们的申请。制药业的检验过程通常包括三个阶段：(1) 临床前检验；(2) 长期使用及安全性检验；(3) 临床效果检验。在每个相继的阶段，药物能通过严格检验的机会都在降低，但是，进一步检验的费用却急剧增加。行业调查表明，研究开发一种新药平均需花费 2.5 亿美元，历时 12 年，因此在检验过程的早期剔除不成功的新药以及识别出有开发前景以供进一步检验的新药就显得极为重要。

统计在药物研究中起到重要作用，因为政府规章非常严格而且执行严厉。在临床前检验中，一般通过两三个典型总体的统计研究来确定一种新药是否可以继续进行长期使用以及安全性检验研究。总体可以由新药、一种控制药物以及一种标准药物组成。当将一种新药送往药理组进行药效评价（药物产生期望效果的能力）时，临床前检验过程就开始了。作为该过程的一部分，需一名统计学家设计一套用于检验新药的试验。该试验必须规定样本容量以及统计分析方法。在两总体研究中，一个样本用于获得有关新药（总体 1）的药效的数据，第二个样本用于获得有关标准药物（总体 2）的药效的数据。根据不同需要，新药及标准药物可能在诸如神经学、心脏病学及免疫学等学科进行检验。在大部分研究中，统计方法涉及新药总体与标准药物总体均值差异的假设检验。如果同标准药物相比，新药缺乏效力或产生了不良的效果，新药就要被拒绝并放弃做进一步检验。只有那些同标准药物相比显示出有前景的新药才会被送去做长期使用及安全性检验。

在药物长期使用和安全性的检测程序中，在临床试验的过程中，要进一步收集数据和进行多总体研究。为避免数据相关的偏差，FDA 要求在进行这些检验前就要确定统计方法。另外，为了避免人为偏差，某些临床试验是双重或三重保密的。也就是说，无论是受试者还是研究者都不知道何种药物分配给谁。如果新药品达到了相关标准药品的所有要求，一份新药的申请（NDA）就会送达 FDA，并由该机构的统计学家及科学家严格地详细审阅。

在本章中，你将学习如何进行两个总体均值和比例的区间估计及假设检验，这些技术是通过分析独立随机样本及匹配样本介绍给大家的。

在第 8 章和第 9 章，我们介绍了对一个总体的均值和比例如何进行区间估计以及假设检验。在本章中，当两个总体的均值之差或比例之差是一个重要的问题时，通过说明如何对两个总体进行区间估计和假设检验，来继续我们有关统计推断的讨论。例如，我们可能要对男女两个总体的平均起薪的差异进行区间估计，或者对供应商 A 与 B 所生产的产品中次品的比例是否存在差异进行假设检验。当假定两个总体的标准差已知时，我们从说明如何进行两个总体均值之差的区间估计和假设检验开始有关统计推断的讨论。

10.1　两总体均值之差的推断：σ_1 和 σ_2 已知

令 μ_1 表示总体 1 的均值，μ_2 表示总体 2 的均值，我们将重点介绍两均值之差：$\mu_1-\mu_2$ 的统计推断。为了进行有关差异的统计推断，我们从总体 1 中抽取一个容量为 n_1 的简单随机样本，从总体 2 中抽取一个容量为 n_2 的另一个简单随机样本。由于这两个样本是相互独立抽取的，因此被称为**独立简单随机样本**（independent simple random samples）。在本节，我们假定可用信息是：在选取样本前两个总体的标准差 σ_1 和 σ_2 可以被假设是已知的，我们称这种情形为 σ_1 和 σ_2 已知的情形。我们在下面例子里说明，当 σ_1 和 σ_2 已知时，如何计算误差范围及如何进行

两个总体均值之差的区间估计。

10.1.1 $\mu_1-\mu_2$的区间估计

Greystone 百货公司在纽约州的布法罗经营两个商场：一个位于市区，另一个地处郊区购物中心。地区经理注意到在一个商场畅销的商品在另一个商场卖得不一定好。经理认为这种情形归因于这两个地区顾客人群的差异。顾客可能在年龄、受教育程度、收入等诸方面存在差异。假定经理要求我们调查这两个商场的顾客平均年龄的差异。

我们定义总体 1 为在市区商场购物的所有顾客，总体 2 为在郊区商场购物的所有顾客。

μ_1 = 总体 1 的均值（即在市区商场购物的所有顾客的平均年龄）

μ_2 = 总体 2 的均值（即在郊区商场购物的所有顾客的平均年龄）

这两个总体的均值差是 $\mu_1-\mu_2$。

为估计 $\mu_1-\mu_2$，我们从总体 1 中抽取一个由 n_1 名顾客组成的简单随机样本，从总体 2 中抽取一个由 n_2 名顾客组成的简单随机样本，然后我们计算这两个样本均值。

$\bar{x}_1$ = n_1 名市区顾客的简单随机样本的样本平均年龄

$\bar{x}_2$ = n_2 名郊区顾客的简单随机样本的样本平均年龄

两总体均值之差的点估计量是两个样本均值之差。

两个总体均值之差的点估计量

$$\bar{x}_1-\bar{x}_2 \tag{10-1}$$

图 10-1 给出了以两个独立简单随机样本为依据，用于估计两个总体均值之差的步骤示意图。

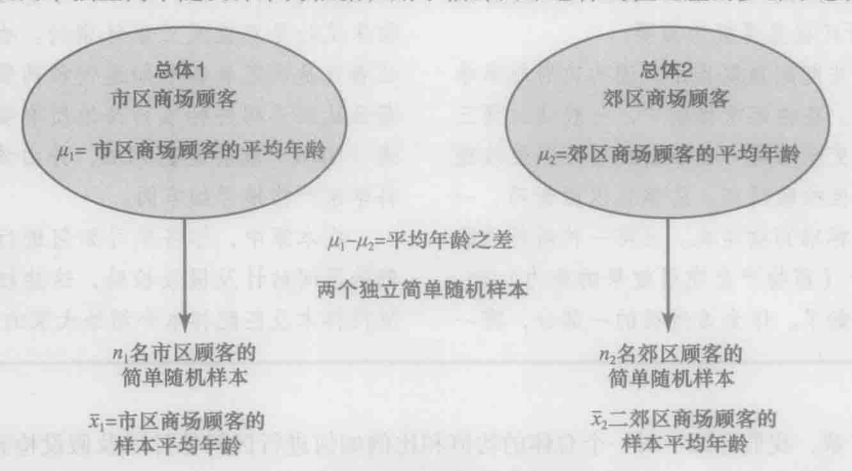


图 10-1 估计两总体的均值之差

与其他点估计量一样，点估计量 $\bar{x}_1-\bar{x}_2$ 有一个描述估计量抽样分布变异性的标准误差。对于两个独立简单随机样本，$\bar{x}_1-\bar{x}_2$ 的标准误差如下[⊖]：

$\bar{x}_1-\bar{x}_2$ 的标准误差

$$\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \tag{10-2}$$

⊖ $\bar{x}_1-\bar{x}_2$ 的标准误差就是 $\bar{x}_1-\bar{x}_2$ 抽样分布的标准差。

如果两个总体都服从正态分布，或者样本容量足够大，使得我们可利用中心极限定理得出：$\bar{x}_1$ 和 $\bar{x}_2$ 的抽样分布近似服从正态分布，那么 $\bar{x}_1 - \bar{x}_2$ 的抽样分布将服从均值为 $\mu_1 - \mu_2$ 的正态分布。

正如我们在第 8 章所介绍的那样，由一个点估计量 ± 一个边际误差就可得出区间估计。在两个总体均值之差估计的例子中，区间估计的公式如下：

$$\bar{x}_1 - \bar{x}_2 \pm \text{边际误差}$$

如果 $\bar{x}_1 - \bar{x}_2$ 的抽样分布服从正态分布，我们可得出边际误差如下：㊀

$$\text{边际误差} = z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2} = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \tag{10-3}$$

于是，两总体均值之差的区间估计如下：

两总体均值之差的区间估计：σ_1 和 σ_2 已知

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \tag{10-4}$$

式中，$1-\alpha$ 为置信系数。

我们再以 Greystone 为例，以先前顾客统计研究的数据为依据，两个总体标准差已知，分别为 $\sigma_1 = 9$ 岁和 $\sigma_2 = 10$ 岁。Greystone 顾客的两个独立简单随机样本收集来的数据给出了下面的结果。

	市区商场	郊区商场
样本容量	$n_1 = 36$	$n_2 = 49$
样本均值	$\bar{x}_1 = 40$ 岁	$\bar{x}_2 = 35$ 岁

利用式（10-1），我们得到两个总体平均年龄之差的一个点估计量为 $\bar{x}_1 - \bar{x}_2 = 40 - 35 = 5$ 岁。于是，我们认为市区商场的顾客平均年龄比郊区商场大 5 岁。现在我们利用式（10-4）来计算边际误差并给出 $\mu_1 - \mu_2$ 的区间估计。在 95% 的置信水平下，$z_{\alpha/2} = z_{0.025} = 1.96$，我们有：

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

$$40 - 35 \pm 1.96\sqrt{\frac{9^2}{36}+\frac{10^2}{49}}$$

$$5 \pm 4.06$$

因此，边际误差为 4.06 岁，两总体均值之差的 95% 的置信区间估计为 $5 - 4.06 = 0.94$ 岁到 $5 + 4.06 = 9.06$ 岁。

10.1.2 $\mu_1 - \mu_2$ 的假设检验

现在我们来讨论两个总体均值之差的假设检验。令 D_0 表示 μ_1 与 μ_2 之间假设的差，假设检验的三种形式如下：

$$H_0: \mu_1 - \mu_2 \geqslant D_0 \qquad H_0: \mu_1 - \mu_2 \leqslant D_0 \qquad H_0: \mu_1 - \mu_2 = D_0$$

$$H_a: \mu_1 - \mu_2 < D_0 \qquad H_a: \mu_1 - \mu_2 > D_0 \qquad H_a: \mu_1 - \mu_2 \neq D_0$$

在许多应用中，$D_0 = 0$，以双侧检验为例，当 $D_0 = 0$ 时，原假设为 $H_0: \mu_1 - \mu_2 = 0$。在这种情形下，原假设为 μ_1 与 μ_2 相等。拒绝 H_0 可推出 $H_a: \mu_1 - \mu_2 \neq 0$ 为真的结论；也就是说，μ_1 与 μ_2 不相等。

第 9 章介绍的假设检验的步骤在这里也是适用的。我们必须选择一个显著性水平，计算检验统计量，并求出 p-值以决定是否应拒绝原假设。对于两个独立简单随机样本，我们发现，式（10-2）给出了点估计量 $\bar{x}_1 - \bar{x}_2$ 的标准误差 $\sigma_{\bar{x}_1-\bar{x}_2}$，并且当样本容量充分大时，$\bar{x}_1 - \bar{x}_2$ 服从正态分布。在这种情形下，当 σ_1 和 σ_2 已知时，两个总体均

㊀ 边际误差等于标准误差乘以 $z_{\alpha/2}$。

值之差的检验统计量如下所示。

$\mu_1-\mu_2$的假设检验的检验统计量：σ_1和σ_2已知

$$z=\frac{(\bar{x}_1-\bar{x}_2)-D_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \tag{10-5}$$

在下面假设检验的例子中，我们将演示如何使用这一检验统计量。

作为评价两个培训中心教育质量差异的研究的一部分，对两个中心的学员进行了一次标准化考试。用考试平均分数的差来评估两个中心教育质量的差异。两个中心的总体均值如下。

μ_1 = 在 A 中心培训的学员总体的平均考试分数

μ_2 = 在 B 中心培训的学员总体的平均考试分数

我们从一个尝试性的假设开始讨论：两个中心的培训质量没有差异。因此，从平均考试分数的角度讲，原假设就是$\mu_1-\mu_2=0$。如果抽样证据导致拒绝该假设，我们就可以得出两总体平均考试分数有差异的结论。该结论表示两个中心教育质量不同并建议对产生这种差异的原因进行跟踪调查研究。对这一双侧检验，原假设和备择假设如下。

$$H_0:\mu_1-\mu_2=0$$
$$H_a:\mu_1-\mu_2\neq 0$$

以前在各种情形下进行的标准化测试总能导致考试分数有近 10 分的标准差。于是，我们将利用这一信息假设总体标准差已知，且$\sigma_1=10$和$\sigma_2=10$。指定研究的显著性水平为$\alpha=0.05$。

从 A 培训中心抽取一个$n_1=30$人的独立简单随机样本，从 B 培训中心抽取一个$n_2=40$人的独立简单随机样本。样本均值分别为$\bar{x}_1=82$、$\bar{x}_2=78$。这些数据是否表明两个培训中心的总体均值存在一个显著差异？为了回答这个问题，我们利用式（10-5）计算检验统计量。

$$z=\frac{(\bar{x}_1-\bar{x}_2)-D_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}=\frac{(82-78)-0}{\sqrt{\frac{10^2}{30}+\frac{10^2}{40}}}=1.66$$

接下来我们计算双侧检验的p-值，由于检验统计量z是在上侧，我们首先计算$z=1.66$的右侧曲线下的面积。利用标准正态分布表，$z=1.66$的左侧曲线下的面积为 0.951 5。因此，该分布上侧的面积为$1.000\,0-0.951\,5=0.048\,5$。由于这个检验是双侧检验，我们必须用 2 乘上侧面积，得到：p-值$=2\times 0.048\,5=0.097\,0$。遵循通常的法则，如果$p$-值$\leqslant\alpha$，则拒绝$H_0$。我们看到，在 0.05 的显著性水平下，$p$-值$=0.097$使我们不能拒绝$H_0$。样本结果不能为我们提供足够的证据来断定两个培训中心的质量是不同的。

本章我们将利用第 9 章所述的p-值法进行假设检验。但是，如果你愿意，也可使用检验统计量和临界值拒绝法则。当$\alpha=0.05$时，$z_{\alpha/2}=z_{0.025}=1.96$。如果$z\leqslant-1.96$或$z\geqslant1.96$，利用临界值法的拒绝法则就会拒绝$H_0$。当$z=1.66$时，我们同样得出不拒绝$H_0$的结论。

在上面的例子中，我们讲述两个总体均值之差的双侧假设检验。我们也可以考虑下侧检验和上侧检验。这些检验使用的检验统计量与式（10-5）相同。对于这些单侧检验，p-值的计算步骤及拒绝法则与第 9 章所述相同。

10.1.3 实用建议

本节所给出的区间估计与假设检验的大部分应用，随机样本都满足$n_1\geqslant30$及$n_2\geqslant30$，一旦其中之一或两者的样本容量都小于 30，总体的分布就需重点加以考虑。在一般情况下，当样本容量较小时，对分析者来说，更为重要的是满足两个总体近似服从正态分布的合理假设。

练习

方法

☆2. 考虑下面的假设检验。

$$H_0: \mu_1 - \mu_2 \leqslant 0 \qquad H_a: \mu_1 - \mu_2 > 0$$

下面的结果来自两个总体的两个独立随机样本。

样本1	样本2
$n_1 = 40$	$n_2 = 50$
$\bar{x}_1 = 25.2$	$\bar{x}_2 = 22.8$
$\sigma_1 = 5.2$	$\sigma_2 = 6.0$

a. 检验统计量的值是多少？

b. p-值是多少？

c. 对 $\alpha = 0.05$，假设检验的结论是什么？

应用

☆4. *Condé Nast Traveler* 杂志进行了一项年度调查，读者对他们最喜爱的豪华游轮进行评分。所有游轮按百分制评分，分值越高表示服务越好。由37艘载客量小于500人的游轮组成一个样本，平均得分为85.36；由44艘载客量大于500人的游轮组成一个样本，平均得分为81.40（*Condé Nast Traveler*，2008年2月）。假定载客量小于500人的游轮的总体标准差为4.55，载客量大于500人的游轮的总体标准差为3.97。

a. 载客量小于500人的游轮和载客量大于500人的游轮的总体平均得分之差的点估计是多少？

b. 对于95%的置信水平，边际误差是多少？

c. 两种载客量游轮的总体平均得分之差的95%的置信区间估计是多少？

6. 假设你负责组织一次商务会议。由于最近经济衰退预算削减，你需要为会议选择一个有最便宜酒店房间的城市。你将选择锁定在亚特兰大和休斯敦两个城市。在本书所附光盘名为Hotel的文件中有亚特兰大和休斯敦酒店房间价格的样本，这些样本数据与Smith旅游研究（Smith Travel Research）报告的结果一致（*SmartMoney*，2009年3月）。因为有大量的关于这两个城市酒店房间价格的历史数据可用，因此亚特兰大房间价格的总体标准差假定为20美元，休斯敦房间价格的总体标准差假定为25美元。基于这些样本数据，你能得出亚特兰大酒店房间的平均价格比休斯敦低的结论吗？

8. 对于提供优质服务的公司，改善客户服务质量将有助于公司股票价格的上涨吗？“研究发现，当一个公司的满意度得分比上年度有所提高，并且高于全美平均水平（目前是75.7）时，从长远来看，该公司的股票有一个优于大盘股市的良好机会”（《商业周刊》，2009年3月2日）。下面是从美国消费者满意度指数得到的三家公司2007年和2008年第四季度的满意度得分。假设满意度得分是以每家公司的60名消费者的投票为依据。由于投票已经进行了好几年，所以在每一种情形下可将三个公司的标准差假定为6分。

公司	2007年得分	2008年得分
Rite Aid	73	76
Expedia	75	77
J. C. Penney	77	78

a. 在 $\alpha = 0.05$ 的显著性水平下，2007～2008年，Rite Aid公司满意度得分的增加在统计上是否显著？你的结论如何？

b. 在 $\alpha = 0.05$ 的显著性水平下，你是否能断定：Rite Aid公司2008年的满意度得分高于75.7的全美平均水平？

c. 在 $\alpha = 0.05$ 的显著性水平下，2007～2008年，Expedia公司满意度得分的增加在统计上是否显著？

d. 在给定标准差、样本容量和 α 数值的情况下进行假设检验时，2007～2008年满意度得分必须增加多少才会在统计上是显著的？

e. 利用（d）的结果来陈述：2007～2008年，J. C. Penney公司满意度得分的增加在统计上是否显著。

10.2 两总体均值之差的推断：σ_1 和 σ_2 未知

在本节，对两总体标准差 σ_1 和 σ_2 未知的情形，我们将延伸有关两总体均值之差统计推断的讨论。在这种情况下，我们用样本标准差 s_1 和 s_2 来估计未知的总体标准差。当我们利用样本标准差时，区间估计与假设检验的程序将建立在 t 分布的基础上而非标准正态分布。

10.2.1 $\mu_1-\mu_2$的区间估计

在下面的例子里我们将介绍，当σ_1和σ_2未知时，如何计算边际误差以及如何建立两总体均值之差的区间估计。Clearwater 国家银行正在进行一项研究，旨在确定其两个分行的客户支票账户余额之间的差异。从 Cherry Grove 分行抽取 28 个支票账户组成一个简单随机样本，从 Beechmont 分行抽取 22 个支票账户组成另一个独立的简单随机样本。每个支票账户的当前余额被记录下来。账户余额汇总如下：

	Cherry Grove	Beechmont
样本容量	$n_1=28$	$n_2=22$
样本均值	$\bar{x}_1=1\,025$ 美元	$\bar{x}_2=910$ 美元
样本标准差	$s_1=150$ 美元	$s_2=125$ 美元

Clearwater 国家银行打算估计 Cherry Grove 和 Beechmont 两个分行客户总体支票账户余额均值的差异。我们来计算这两个总体均值之差的边际误差和区间估计。

在第 10.1 节，当标准差σ_1和σ_2已知时，我们给出了下面的区间估计。

$$\bar{x}_1-\bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

当σ_1和σ_2未知时，我们用样本标准差s_1和s_2来估计σ_1和σ_2，并用$t_{\alpha/2}$代替$z_{\alpha/2}$。则两总体均值之差的区间估计表达式如下。

两个总体均值之差的区间估计：σ_1和σ_2未知

$$\bar{x}_1-\bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \tag{10-6}$$

式中，$1-\alpha$为置信系数。

在该式中，所用的t分布是近似分布，但估计结果很好且应用相对简单。在应用式（10-6）的过程中遇到的唯一难题是确定$t_{\alpha/2}$合适的自由度。统计软件包可自动计算合适的自由度。所用公式如下：[⊖]

自由度：两个独立随机样本的t分布

$$df=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \tag{10-7}$$

我们继续以 Clearwater 国家银行为例，来介绍如何应用式（10-6）来估计两个分行支票账户余额总体均值之差的95%的置信区间估计。Cherry Grove 分行的样本数据是$n_1=28$，$\bar{x}_1=1\,025$美元，$s_1=150$美元；Beechmont 分行的样本数据是$n_2=22$，$\bar{x}_2=910$美元，$s_2=125$美元。$t_{\alpha/2}$的自由度计算如下

$$df=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}=\frac{\left(\frac{150^2}{28}+\frac{125^2}{22}\right)^2}{\frac{1}{28-1}\left(\frac{150^2}{28}\right)^2+\frac{1}{22-1}\left(\frac{125^2}{22}\right)^2}=47.8$$

为得到一个较大的t值和更保守的区间估计，我们把非整的自由度取整为47。查自由度为47的t分布表，可得$t_{0.025}=2.012$。利用式（10-6），我们建立两总体均值之差的95%的置信区间如下。

⊖ 当用s_1和s_2来估计σ_1和σ_2时，t分布用于对两总体均值之差进行统计推断。

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$1\,025 - 910 \pm 2.012\sqrt{\frac{150^2}{28} + \frac{125^2}{22}}$$

$$115 \pm 78$$

两分行支票账户余额总体均值之差的点估计值为 115 美元，边际误差为 78 美元，两总体均值之差的 95% 的置信区间估计为 115 - 78 = 37 美元到 115 + 78 = 193 美元。

若利用式（10-7）人工计算自由度是非常繁杂的，而用计算机软件包就很容易了。注意：在式（10-6）和式（10-7）中都有 s_1^2/n_1 和 s_2^2/n_2，为了估计式（10-6）和式（10-7），这些值仅需要计算一次。[⊖]

10.2.2　$\mu_1 - \mu_2$ 的假设检验

现在我们来讨论总体标准差 σ_1 和 σ_2 未知时的两总体均值之差的假设检验。假设 μ_1 与 μ_2 之差可用 D_0 表示，在第 10.1 节，当 σ_1 和 σ_2 已知时，我们给出了下面的检验统计量。

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

检验统计量 z 服从标准正态分布。

当 σ_1 和 σ_2 未知时，我们用 s_1 作为 σ_1 的估计量，s_2 作为 σ_2 的估计量。当 σ_1 和 σ_2 未知时，我们用这两个样本标准差替代 σ_1 和 σ_2，可得以下的检验统计量。

$\mu_1 - \mu_2$ 的假设检验的检验统计量：σ_1 和 σ_2 未知

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{10-8}$$

t 统计量的自由度由式（10-7）给出。

在下面的假设检验例子中，我们将说明这一检验统计量的用法。

设想开发一个新的计算机软件包，它有助于系统分析员减少设计、开发、实现信息系统所需要的时间。为评估新软件包的优点，抽取了 24 名系统分析员组成一个随机样本。发给每个分析员一张假定的信息系统的说明书。指定其中 12 名分析员使用当前技术来开发该信息系统，另外 12 名分析员先接受新软件包的培训，然后用新软件包来开发该信息系统。

在本研究中有两个总体：使用当前技术的系统分析员总体和使用新软件包的系统分析员总体。考虑完成该信息系统设计项目所需要的时间，总体均值如下。

μ_1——使用当前技术的系统分析员平均完成项目时间

μ_2——使用新软件包的系统分析员平均完成项目时间

负责评估新软件的研究人员希望证明新软件包能使平均完成项目时间缩短。因此，研究人员要寻找证据来验证 μ_2 是小于 μ_1 的；在这种情形下，两总体均值之差 $\mu_1 - \mu_2$ 将大于 0。研究的假设 $\mu_1 - \mu_2 > 0$ 是作为备择假设提出来的。因此，假设检验为

$$\mathrm{H}_0: \mu_1 - \mu_2 \leq 0$$

$$\mathrm{H}_a: \mu_1 - \mu_2 > 0$$

我们使用的显著性水平为 $\alpha = 0.05$。

⊖　如果你利用式（10-7）手工计算自由度，该建议应该是有帮助的。

假定24个分析员完成研究的时间如表10-1所示，利用式（10-8）给出的检验统计量，我们得到

$$t=\frac{(\bar{x}_1-\bar{x}_2)-D_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}=\frac{(325-286)-0}{\sqrt{\frac{40^2}{12}+\frac{44^2}{12}}}=2.27$$

表10-1 软件检验研究的完成时间数据与汇总统计

	当前技术	新软件
	300	274
	280	220
	344	308
	385	336
	372	198
	360	300
	288	315
	321	258
	376	318
	290	310
	301	332
	283	263
汇总统计量		
样本容量	$n_1=12$	$n_2=12$
样本均值	$\bar{x}_1=325$（小时）	$\bar{x}_2=286$（小时）
样本标准差	$s_1=40$	$s_2=44$

利用式（10-7）计算自由度，我们得到

$$df=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}=\frac{\left(\frac{40^2}{12}+\frac{44^2}{12}\right)^2}{\frac{1}{12-1}\left(\frac{40^2}{12}\right)^2+\frac{1}{12-1}\left(\frac{44^2}{12}\right)^2}=21.8$$

舍入后，我们将使用自由度为21的 t 分布表。t 分布表的这一行如下所示：

上侧面积	0.20	0.10	0.05	0.025	0.01	0.005
t 值（自由度21）	0.859	1.323	1.721	2.080	2.518	2.831

↑ $t=2.27$

对于上侧检验，p-值就是 $t=2.27$ 右侧曲线下的面积。从上面结果可知，p-值处于0.025与0.01之间。因此，p-值小于 $\alpha=0.05$，并且 H_0 被拒绝。样本结果能使研究者得出结论：$\mu_1-\mu_2>0$，或 $\mu_1>\mu_2$。于是，调查研究证实了新的软件包能减少平均完成项目的时间。[⊖]

Minitab或Excel能用来分析两总体均值之差的检验假设的数据。如图10-2所示，Minitab的输出结果比较了当前的和新的软件技术。输出的最后一行表示 $t=2.27$ 及 p-值$=0.017$。注意，Minitab利用了式（10-7）计算出这个分析的自由度是21。

```
Two-sample T for Current vs New

          N    Mean    StDev   SE Mean
Current  12   325.0    40.0     12
New      12   286.0    44.0     13

Difference = mu Current - mu New
Estimate for difference:  39.0000
95% lower bound for difference = 9.5
T-Test of difference = 0 (vs >):  T-Value = 2.27  P-Value = 0.017  DF = 21
```

图10-2 当前的和新的软件技术假设检验的MINITAB输出结果

10.2.3 实用建议

本节介绍的区间估计与假设检验方法是一种强有力的实用方法，并且可用于相对较小的样本容量。在大部分应用中，如果两个总体的样本容量 n_1 和 n_2 相等或接近相等，使得总样本容量 n_1+n_2 至少为20时，即使总体不是

⊖ 应用 t 分布表，我们只能确定 p-值的一个范围，利用Excel或Minitab，表明精确的 p-值$=0.017$。

正态分布，我们也能期望得到非常好的结果。如果总体分布高度偏斜或含有异常点，我们建议使用较大的样本容量。只有当分析者认为总体分布近似服从正态分布时才可使用较小的样本容量。[⊖]

注释和评论

当σ_1和σ_2未知时，另一种对两总体均值之差进行统计推断的方法，是以假设两总体的标准差相等（$\sigma_1=\sigma_2=\sigma$）为依据。在此假设下，将两个样本方差组合起来，得到合并的样本方差：

$$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

检验统计量t变为

$$t=\frac{(\bar{x}_1-\bar{x}_2)-D_0}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$$

并且自由度为n_1+n_2-2。在这里，p-值的计算及样本结果的解释与本节较早时讨论的步骤相同。

这个过程中的一个难题就是两总体标准差相等的假设通常难以验证，我们经常遇到的是不等的标准差。尤其当样本容量n_1和n_2相差很大时，应用合并步骤不可能提供满意的结果。

本节介绍的t检验方法并不需要总体标准差相等的假设，并且不论总体标准差相等与否都是适用的。在大部分的应用中，t检验是被我们推荐使用的较为一般的方法。

练习

方法

☆10. 考虑下面的假设检验。

$$H_0: \mu_1-\mu_2=0$$
$$H_a: \mu_1-\mu_2\neq 0$$

下面的结果来自两个总体的两个独立随机样本。

样本1	样本2
$n_1=35$	$n_2=40$
$\bar{x}_1=13.6$	$\bar{x}_2=10.1$
$s_1=5.2$	$s_2=8.5$

a. 检验统计量的值是多少？

b. t分布的自由度是多少？

c. p-值是多少？

d. 若$\alpha=0.05$，你的结论是什么？

应用

☆12. 美国交通部（U. S. Department of Transportation）给出了75个最大的大都会区居民每天一辆汽车行驶的英里数。假设由50名布法罗居民组成的简单随机样本，平均每天行驶22.5英里，标准差为每天8.4英里；由40名波士顿居民组成的独立简单随机样本，平均每天行驶18.6英里，标准差为每天7.4英里。

a. 布法罗居民和波士顿居民每天平均行驶的英里数之差的点估计是多少？

b. 两个总体均值之差的95%的置信区间是多少？

14. 佛罗里达州坦帕市护士的薪金水平比得克萨斯州达拉斯市低吗？薪金数据显示，坦帕医院护士的薪金要比达拉斯的低（*The Tampa Tribune*，2007年1月15日）。假设对坦帕的40名医院护士和达拉斯的50名医院护士进行一项后续研究，得到如下结果。

坦帕	达拉斯
$n_1=40$	$n_2=50$
$\bar{x}_1=56\,100$ 美元	$\bar{x}_2=59\,400$ 美元
$s_1=6\,000$ 美元	$s_2=7\,000$ 美元

a. 提出这样的假设，如果原假设被拒绝，我们可以得到结论：坦帕医院护士的薪金显著的低于达拉斯医院护士的薪金。显著性水平为$\alpha=0.05$。

b. 检验统计量的值是多少。

c. p-值是多少。

d. 你的结论是什么？

⊖ 无论是否可能，相同的样本容量$n_1=n_2$都是可取的。

16. 美国大学理事会以学生父母所受的最高教育水平为依据，比较了 SAT 的分数。一项研究的假设是，父母受教育水平较高的学生，其 SAT 的平均分数也比较高。SAT 数学分数的总体平均分是 514 分（College Board website，2012 年 1 月 8 日）。对于独立的学生样本，SAT 数学分数如下表所示。第一个样本是父母为大学毕业且拥有学士学位的学生的 SAT 数学分数。第二个样本是父母为高中毕业没有大学学位的学生的 SAT 数学分数。

学生父母			
大学毕业		高中毕业	
485	487	442	492
534	533	580	478
650	526	479	425
554	410	486	485
550	515	528	390
572	578	524	535
497	448		
592	469		

a. 提出可以用来确定样本数据是否支持的假设：如果父母受教育程度较高，则学生数学总体平均分数也比较高。

b. 两总体平均分数之差的点估计是多少？

c. 计算假设检验的 p-值。

d. 在 $\alpha=0.05$ 的显著性水平下，你的结论是什么？

18. 美国普渡大学和威奇塔州立大学的研究人员发现，航空公司正在做一件有益的工作：让乘客能准点到达目的地（美联社，2012 年 4 月 2 日）。美国穿越航空公司（AirTran Airways）和西南航空公司（Southwest Airlines）的航班，有 88% 能准点到达目的地，从而跻身准点到达的航空公司之中。但是对于 12% 延误的航班，这些航班延误的时间是多少分钟呢？在本书所附光盘名为 AirDelay 的文件中，给出了这两个航空公司的航班延误时间（单位：分）的样本数据。

a. 建立能用来检验这两个航空公司延误航班的总体平均延误时间之差的假设。

b. 每个航空公司延误航班的样本平均延误时间是多少？

c. 在 $\alpha=0.05$ 的显著性水平下，p-值是多少？你的结论是什么？

10.3 两总体均值之差的推断：匹配样本

假设一家制造企业的员工可用两种不同方法完成一项生产任务。为了使产品产量最大化，公司想确认总体完成生产任务平均时间较少的方法。令 μ_1 表示生产方法 1 的总体完成生产任务的平均时间，μ_2 表示生产方法 2 的总体完成生产任务的平均时间。在不知道哪种方法更好的条件下，我们暂时假设两种方法具有相同的平均完成生产任务时间。因此，原假设为：$H_0:\mu_1-\mu_2=0$。如果拒绝原假设，我们就可得出总体完成生产任务的平均时间不等的结论。在这种情况下，就可找出完成生产任务平均时间较少的方法。原假设和备择假设如下。

$$H_0:\mu_1-\mu_2=0$$
$$H_a:\mu_1-\mu_2\neq 0$$

在选择用于搜集生产时间数据及检验假设的抽样方法时，我们考虑两种选择方案。一种是基于**独立样本**（independent sample），另一种是基于**匹配样本**（matched samples）。

1. 独立样本设计：抽取工人的一个简单随机样本，样本中的每个工人使用生产方法 1；抽取工人的另一个独立的简单随机样本，样本中的每个工人使用生产方法 2。总体均值差的检验使用第 10.2 节的方法。

2. 匹配样本设计：抽取工人的一个简单随机样本，每个工人先用一种生产方法，然后用另一种生产方法。两种方法的次序被随机地指派给工人，一些工人先使用生产方法 1，其他工人先使用生产方法 2。每个工人提供一对数据值，一个数值是生产方法 1 的，另一个数值是生产方法 2 的。

在匹配样本设计中，两种生产方法在相似条件下被检验（即由相同工人使用）；因此这一设计产生的抽样误差往往比独立样本设计要小。这主要是因为在匹配样本设计中，两种生产方法被相同的工人使用，剔除了工人之间的差异。

为了演示分析匹配样本设计，现在我们利用匹配样本设计对两种生产方法总体均值之差进行检验。现有 6 名工人组成一个随机样本。6 名工人完成生产任务时间的数据由表 10-2 给出。注意：每名工人提供一对数据值，每

个数据值对应一种生产方法。我们还注意到，最后一列给出了样本中每名工人完成生产任务的时间之差 d_i。

分析匹配样本设计的关键是要意识到，我们仅考虑差值这一列。因此，我们有 6 个数据值（0.6，-0.2，0.5，0.3，0.0 和 0.6）可用于分析两种生产方法总体均值之差。

表 10-2 匹配样本设计的任务完成时间

工人	方法 1 的完成时间（分）	方法 2 的完成时间（分）	完成时间的差（d_i）
1	6.0	5.4	0.6
2	5.0	5.2	-0.2
3	7.0	6.5	0.5
4	6.2	5.9	0.3
5	6.0	6.0	0.0
6	6.4	5.8	0.6

令 μ_d 表示工人总体数值之差的平均值。利用该符号，重新写出下面的原假设和备择假设。

$$H_0: \mu_d = 0$$
$$H_a: \mu_d \neq 0$$

如果 H_0 被拒绝，我们可以得出总体完成生产任务的平均时间是不同的结论。

符号 d 提示我们：匹配样本给出的是差值的数据。表 10-2 中 6 个差值的样本均值与样本标准差如下。[㊀]

$$\bar{d} = \frac{\sum d_i}{n} = \frac{1.8}{6} = 0.30$$

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{0.56}{5}} = 0.335$$

对于 $n=6$ 名工人的小样本，我们需要假设差值的总体服从正态分布。该假设对于运用 t 分布进行假设检验和区间估计是必要的。下面是自由度为 $n-1$ 的 t 分布的检验统计量。[㊁]

匹配样本假设检验的检验统计量

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} \tag{10-9}$$

在 $\alpha=0.05$ 的显著性水平下，我们用式（10-9）来检验假设 $H_0: \mu_d=0$ 和 $H_a: \mu_d \neq 0$。将样本值 $\bar{d}=0.30$，$s_d=0.335$ 及 $n=6$ 代入式（10-9），我们可以计算出检验统计量的值。[㊂]

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} = \frac{0.30 - 0}{0.335 / \sqrt{6}} = 2.20$$

现在我们计算双侧检验的 p-值。由于 $t=2.20>0$，所以检验统计量位于 t 分布的上侧。因为 $t=2.20$，利用自由度 $=n-1=6-1=5$ 的 t 分布表，可以得到检验统计量右侧曲线下的面积。自由度为 5 的 t 分布表的信息如下：

上侧面积	0.2	0.1	0.05	0.025	0.01	0.005
t 值（自由度为 5）	0.920	1.476	2.015	2.571	3.365	4.032

↑ $t=2.20$

因此，我们看到：右侧曲线下的面积介于 0.05 到 0.025 之间。由于该检验为双侧检验，我们应将 0.05 和 0.025 增大一倍，从而得到 p-值介于 0.10～0.05 的结论。p-值大于 $\alpha=0.05$，因此原假设 $H_0: \mu_d=0$ 未被拒绝。利用 Excel 或 Minitab 和表 10-2 的数据，我们得到精确的 p-值 $=0.080$。

另外，利用第 8 章一个总体的方法，我们能够得到两总体均值之差的区间估计。对于 95% 的置信水平，计算结果如下。

㊀ 除了符号 d 的使用外，样本均值与样本标准差的公式与本节前面用到的一样。
㊁ 如果样本容量足够大，总体服从正态分布的假设就不必要了。在第 8 章和第 9 章中已经介绍了使用 t 分布的样本容量准则。
㊂ 一旦计算出差值数据，匹配样本的 t 分布的步骤就与第 8 章和第 9 章中描述的一个总体的估计和假设检验步骤一样。

$$\bar{d} \pm t_{0.025}\frac{s_d}{\sqrt{n}}$$

$$0.30 \pm 2.571 \times \frac{0.335}{\sqrt{6}}$$

$$0.30 \pm 0.35$$

于是，边际误差为0.35，两种生产方法总体均值之差的95%的置信区间为-0.05~0.65分钟。

注释和评论

1. 在本节介绍的例子中，工人们首先用第一种生产方法完成生产任务，然后再用第二种生产方法。这个例子说明了在匹配样本设计中每个抽样元素（工人）提供了一对数据值。使用不同的，但"类似的"元素也能提供成对的数据值。例如，某地的一个工人可能与另一地的一个类似的工人匹配（基于年龄、教育、性别、经验等方面的类似）。这些成对的工人可以提供用于匹配样本分析的差值数据。
2. 用匹配样本设计方法进行两总体均值的统计推断一般比用独立样本方案更精确，因此是被推荐的设计。然而在某些应用中，这种匹配不能完成，或是与匹配相关的时间过长和费用过高，在这种情形下，还是应当使用独立样本设计方案。

练习

方法

20. 下面的数据取自两个总体的匹配样本。

元素	总体 1	总体 2
1	11	8
2	7	8
3	9	6
4	12	7
5	13	10
6	15	15
7	15	14

a. 计算每个元素的差值。

b. 计算 $\bar{d}$。

c. 计算标准差 s_d。

d. 两总体均值之差的点估计是多少？

e. 建立一个两总体均值之差的95%的置信区间。

应用

22. 由25家公司组成了一个随机样本，该样本记录了2012年年初和2012年第1季度末的每股股价（《华尔街日报》，2012年4月2日）。在第一季度，有多少股票起到了未来股市和经济指标的作用。利用存放在本书所附光盘名为StockPrices的样本数据文件回答以下问题。

a. 设 d_i 表示第 i 家公司每股股价的变化，其中 d_i 是2012年第1季度末的每股股价与2012年年初的每股股价之差。利用这些差值的样本均值，估计每股股价在第1季度变化的金额（单位：美元）。

b. 在第一季度每股股价总体平均变化的95%的的置信区间估计是多少？

24. 全球商务旅行协会报告了本年度及上一年度国内商务旅行的机票价格（*INC. Magazine*, 2012年2月）。下表是这两个年度的12个航班及其国内机票的样本资料。

本年度	上一年度
345	315
526	463
420	462
216	206
285	275
405	432
635	585
710	650
605	545
517	547
570	508
610	580

a. 提出假设并且检验：一年期的国内商务旅行机票的平均价格是否有了一个显著增长。p-值是多少？在 $\alpha = 0.05$ 的显著性水平下，你的结论是什么？

b. 每个年度的国内商务旅行机票的样本平均价格是多少？

c. 一年期的国内商务旅行机票价格的百分比变化是多少？

26. 由20名参加美国职业高尔夫球协会（PGA）比赛的高尔夫球员组成一个样本，他们在第一轮和第四轮（最后一轮）的杆数如下表所示（*Golfweek*, 2009年2月14日和2009年2月28日）。假设你想确定，对于一次PGA巡回赛，第一轮的平均杆数是否与第四轮（最后一轮）的平均杆数显著不同。最后一轮的比赛压力会使杆数上升吗？或者，增加高尔夫球员的注意力反而会使杆数下降？

球员	第一轮	第四轮
Michael Letzig	70	72
Scott Verplank	71	72
D. A. Points	70	75
Jerry Kelly	72	71
Soren Hansen	70	69
D. J. Trahan	67	67
Bubba Watson	71	67
Reteif Goosen	68	75
Jeff Klauk	67	73
Kenny Perry	70	69
Aron Price	72	72
Charles Howell	72	70
Jason Dufner	70	73
Mike Weir	70	77
Carl Pettersson	68	70
Bo Van Pelt	68	65
Ernie Els	71	70
Cameron	70	68
Nick Watney	69	68
Tommy Armour	67	71

a. 在 $\alpha=0.10$ 的显著性水平下，检验第一轮和第四轮的总体平均杆数在统计上是否存显著差异。p-值是多少？你的结论是什么？

b. 两总体均值之差的点估计是多少？哪一轮的总体平均杆数较少？

c. 两总体均值之差的90%的置信区间估计的边际误差是多少？这一置信区间能用于检验（a）中的假设吗？请作出解释。

10.4 两总体比例之差的推断

令 p_1 表示总体1的比例，p_2 表示总体2的比例，下面我们来讨论两总体比例之差 p_1-p_2 的统计推断。为了对这个比例之差作出推断，我们选择两个独立的随机样本，这两个样本分别由总体1中的 n_1 个单位和总体2中的 n_2 个单位组成。

10.4.1 p_1-p_2 的区间估计

在下面的例子中，我们将介绍如何计算边际误差以及如何建立两总体比例之差的区间估计。

一家报税公司想要比较该公司两个地区办事处的工作质量。在每个办事处，随机抽取准备报税的申报单样本，并对申报单的准确性进行核实，该报税公司可以估计出每个办事处准备报税的申报单的出错比例。特别值得注意的是这些比例之差。

p_1——总体1（办事处1）的出错申报单的比例

p_2——总体2（办事处2）的出错申报单的比例

$\bar{p}_1$——总体1的简单随机样本的样本比例

$\bar{p}_2$——总体2的简单随机样本的样本比例

两总体比例之差为 p_1-p_2。p_1-p_2 的点估计量如下。

两总体比例之差的点估计量

$$\bar{p}_1-\bar{p}_2 \tag{10-10}$$

于是，两总体比例之差的点估计量是两个独立的简单随机样本的样本比例之差。

与其他的点估计量一样，如果我们重复地抽取两个独立的随机样本，则点估计量 $\bar{p}_1-\bar{p}_2$ 有一个反映它所有可能值的抽样分布。这个抽样分布的均值为 p_1-p_2，标准误差如下：

$\bar{p}_1-\bar{p}_2$ 的标准误差

$$\sigma_{\bar{p}_1-\bar{p}_2}=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}} \tag{10-11}$$

如果样本容量足够大，使得 n_1p_1，$n_1(1-p_1)$，n_2p_2 和 $n_2(1-p_2)$ 都大于或等于5，则 $\bar{p}_1-\bar{p}_2$ 的抽样分布近似服从正态分布。

如前所述，区间估计可由点估计加减边际误差给出。在两总体比例之差的估计中，区间估计将采取以下形式：

$$\bar{p}_1-\bar{p}_2\pm\text{边际误差}$$

由于 $\bar{p}_1-\bar{p}_2$ 的抽样分布近似服从正态分布，我们可用 $z_{\alpha/2}\sigma_{\bar{p}_1-\bar{p}_2}$ 作为边际误差。但是，因为两总体比例 p_1 和 p_2 未知，所以式（10-11）给出的 $\sigma_{\bar{p}_1-\bar{p}_2}$ 不能直接使用。我们用样本比例 $\bar{p}_1$、$\bar{p}_2$ 来估计 p_1 和 p_2，得到的边际误差如下。

$$\text{边际误差}=z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}+\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \tag{10-12}$$

两总体比例之差区间估计的一般形式如下。

两总体比例之差的区间估计

$$\bar{p}_1-\bar{p}_2\pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}+\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \tag{10-13}$$

式中，$1-\alpha$ 为置信系数。

回到报税公司的例子，我们得到两个办事处的独立简单随机样本的信息如下。

办事处1	办事处2
$n_1=250$	$n_2=300$
出错申报单数 = 35	出错申报单数 = 27

两个办事处的样本比例如下。

$$\bar{p}_1=\frac{35}{250}=0.14$$

$$\bar{p}_2=\frac{27}{300}=0.09$$

两总体出错申报单比例之差的点估计为 $\bar{p}_1-\bar{p}_2=0.14-0.09=0.05$。于是我们估计出办事处1的出错率比办事处2多0.05或5%。

现在利用式（10-13）计算边际误差及两总体比例之差的区间估计。对于90%置信区间，$z_{\alpha/2}=z_{0.05}=1.645$，于是有

$$\bar{p}_1-\bar{p}_2\pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}+\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

$$0.14-0.09\pm1.645\sqrt{\frac{0.14\times(1-0.14)}{250}+\frac{0.09\times(1-0.09)}{300}}$$

$$0.05 \pm 0.045$$

于是，边际误差为0.045，90%的置信区间为（0.005，0.095）。

10.4.2 $p_1 - p_2$的假设检验

现在我们考虑两总体比例之差的假设检验。我们把重点放在两总体比例无差异的检验上。在这种情况下，假设检验有如下三种形式：[⊖]

$$H_0: p_1 - p_2 \geqslant 0 \qquad H_0: p_1 - p_2 \leqslant 0 \qquad H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 < 0 \qquad H_a: p_1 - p_2 > 0 \qquad H_a: p_1 - p_2 \neq 0$$

当我们假定作为一个等式的原假设H_0为真时，有$p_1 - p_2 = 0$，也就是说两总体比例是相等的，即$p_1 = p_2$。

我们将以点估计量$\bar{p}_1 - \bar{p}_2$的抽样分布作为检验统计量的基础。在式（10-11）中，我们已经给出$\bar{p}_1 - \bar{p}_2$的标准误差为

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

在假定作为一个等式的H_0为真的情况下，总体比例相等且$p_1 = p_2 = p$。这时，$\sigma_{\bar{p}_1 - \bar{p}_2}$变为

当$p_1 = p_2 = p$时，$\bar{p}_1 - \bar{p}_2$的标准误差

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad (10\text{-}14)$$

由于p未知，我们合并或组合两个样本点估计量（$\bar{p}_1$和$\bar{p}_2$），得到p的单一的点估计量如下：

当$p_1 = p_2 = p$时，p的合并估计量

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} \qquad (10\text{-}15)$$

这个p的**合并估计量**（pooled estimator of p）是$\bar{p}_1$和$\bar{p}_2$的加权平均值。

用$\bar{p}$代替式（10-14）中的p，我们可以得到$\bar{p}_1 - \bar{p}_2$的标准误差的估计。检验统计量使用标准误差的这个估计。关于两总体比例之差的假设检验的检验统计量，其一般形式是点估计量除以$\sigma_{\bar{p}_1 - \bar{p}_2}$的一个估计。

$p_1 - p_2$的假设检验的检验统计量

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (10-16)$$

该检验统计量适用于n_1p_1、$n_1(1-p_1)$、n_2p_2和$n_2(1-p_2)$全都大于或等于5的大样本情况。

让我们仍以报税公司为例，并假设该公司希望应用假设检验来确定两个办事处之间的出错比例是否不同。双侧检验的原假设和备择假设如下：

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 \neq 0$$

如果H_0被拒绝，该公司就可以得出两个办事处出错率不同的结论。我们的显著性水平为$\alpha = 0.10$。

根据先前收集的样本数据，从办事处1抽取$n_1 = 250$份申报单，得到$\bar{p}_1 = 0.14$；从办事处2抽取$n_2 = 300$份申报单，得到$\bar{p}_2 = 0.09$。接着计算p的合并估计量。

⊖ 所有假设我们都考虑用0作为两个被关注量之间的差。

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = \frac{250 \times 0.14 + 300 \times 0.09}{250 + 300} = 0.1127$$

利用这个合并估计量和样本比例之差，检验统计量的值如下。

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.14 - 0.09}{\sqrt{0.1127 \times (1 - 0.1127) \times \left(\frac{1}{250}+\frac{1}{300}\right)}} = 1.85$$

计算这一双侧检验的p-值，我们首先注意到：$z=1.85$位于标准正态分布的上侧。利用标准正态分布表，我们发现$z=1.85$右侧曲线下的面积为$1.0000-0.9678=0.0322$。对于双侧检验，将这个面积乘以2，我们得到p-值$=2\times0.0322=0.0644$。由于p-值小于$\alpha=0.10$，在0.10的显著性水平下我们拒绝H_0。于是，报税公司可以得出两个办事处出错率不同的结论。这个假设检验的结论与前面两个办事处总体出错率之差的区间估计为（0.005，0.095）是一致的，办事处1的出错率更高一些。

练 习

方法

☆28. 考虑下面的取自两个总体的两个独立样本的数据。

样本1	样本2
$n_1=400$	$n_2=300$
$\bar{p}_1=0.48$	$\bar{p}_2=0.36$

a. 两总体比例之差的点估计是多少？

b. 建立两总体比例之差的90%的置信区间。

c. 建立两总体比例之差的95%的置信区间。

应用

30.《商业周刊》/哈里斯调查就未来的经济前景向大公司的一些高级经理进行了一次问卷调查。一个问题是“你认为你的公司在未来12个月内，全日制雇员的人数会增加吗?”在这次调查中，400名经理中有220人回答“是”，而在去年的调查中，400名经理中有192人回答“是”。建立这两个时点比例之差的95%的置信区间估计。你对该区间估计的解释是什么？

32. 美国汽车协会（AAA）的一项研究调查了男性还是女性更可能停车问路的问题（AAA，2006年1月）。研究中涉及以下问题：“如果你和你的配偶一起驾车出行迷路了，你会停下来问路吗?”AAA采用的一份具有代表性的样本数据显示，811名女性中有300名回答她们会停车问路，而750名男性中有255名回答他们会停车问路。

a. AAA研究的假设是：女性更可能回答她们会停车问路。建立这一研究的原假设和备择假设。

b. 愿意停车问路的女性的比例是多少？

c. 愿意停车问路的男性的比例是多少？

d. 在$\alpha=0.05$的显著性水平下，检验该假设。p-值是多少？你预计AAA从这项研究中会得出什么结论？

34.《商业周刊》报道，不同年龄组的俄罗斯人对生活在俄罗斯的喜欢程度似乎不同（《商业周刊》，2008年3月10日）。下面的样本数据与《商业周刊》的调查结果是一致的，这些数据反映了不同年龄组的人对“你是否喜欢生活在俄罗斯”这一问题的回答。

	俄罗斯人的年龄组	
	17~26岁	40岁以上
样本	300	260
回答“是”	192	117

a. 17~26岁的俄罗斯人，喜欢生活在俄罗斯的总体比例的点估计是多少？

b. 40岁以上的俄罗斯人，喜欢生活在俄罗斯的总体比例的点估计是多少？

c. 建立17~26岁年轻的俄罗斯人和40岁以上年长的俄罗斯人，喜欢生活在俄罗斯的总体比例之差的95%的置信区间估计。

36. 冬季游客对于佛罗里达州西南部地区的经济非常重要。酒店入住率是一个经常被报道的计量游客数量和游客活动的依据（*Naples Daily News*，2012年3月22日）。连续两年2月份的酒店入住数据如下表所示。

	本年度	上年度
入住房间	1470	1458
总房间	1750	1800

a. 提出能用来确定：在这连续的两年里，酒店入住率是否有了显著增长的假设。

b. 每年酒店入住率的点估计是多少？

c. 在 $\alpha=0.05$ 的显著性水平下，你假设检验的结论是什么？p-值是多少？

d. 在这连续的两年里，酒店入住率变化的95%的置信区间估计。你认为，该地区的官员对这一结果会感到很满意吗？

小结

本章我们讨论了涉及两个总体建立区间估计及进行假设检验的方法。首先，我们阐述了在抽取简单独立随机样本的情况下，如何进行关于两个总体均值之差的推断。我们首先考虑了假定总体标准差 σ_1 和 σ_2 已知的情况。标准正态分布 z 可用来建立区间估计和假设检验的检验统计量。然后，我们考虑了总体标准差未知并由样本标准差 s_1 和 s_2 估计的情形。在这种情况下，可用 t 分布来进行区间估计和作为假设检验的检验统计量。

接下来我们讨论了匹配样本设计的两总体均值之差的统计推断。在匹配样本设计中，每个元素提供一对数据值，每一个总体有一个。然后在统计分析中利用这些成对数据值之差。因为在大多数情形下，匹配样本方法会提高估计的精确度，因此匹配样本设计通常优于独立样本设计。

最后，讨论了关于两个总体比例之差的区间估计与假设检验。用于分析两个总体比例之差的统计方法类似于分析两个总体均值之差的统计方法。

关键术语

independent sample 独立样本 取自两个总体的样本，抽取的方式是：组成一个样本的元素与组成另一个样本的元素是独立选取的。

matched sample 匹配样本 一个样本的每个数据值与另一个样本的对应数据值相匹配的样本。

pooled estimator of p p的合并估计量 总体比例的一个估计量，它是两个独立样本的点估计量的加权平均数。

重要公式

两个总体均值之差的点估计量

$$\bar{x}_1-\bar{x}_2 \tag{10-1}$$

$\bar{x}_1-\bar{x}_2$ 的标准误差

$$\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \tag{10-2}$$

两个总体均值之差的区间估计：σ_1 和 σ_2 已知

$$\bar{x}_1-\bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \tag{10-4}$$

$\mu_1-\mu_2$ 的假设检验的检验统计量：σ_1 和 σ_2 已知

$$z=\frac{(\bar{x}_1-\bar{x}_2)-D_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \tag{10-5}$$

两个总体均值之差的区间估计：σ_1 和 σ_2 未知

$$\bar{x}_1-\bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \tag{10-6}$$

利用两独立随机样本的 t 分布的自由度

$$df=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \tag{10-7}$$

$\mu_1-\mu_2$ 的假设检验的检验统计量：σ_1 和 σ_2 未知

$$t=\frac{(\bar{x}_1-\bar{x}_2)-D_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \tag{10-8}$$

匹配样本假设检验的检验统计量

$$t=\frac{\bar{d}-\mu_d}{s_d/\sqrt{n}} \tag{10-9}$$

两总体比例之差的点估计量

$$\bar{p}_1-\bar{p}_2 \tag{10-10}$$

$\bar{p}_1-\bar{p}_2$ 的标准误差

$$\sigma_{\bar{p}_1-\bar{p}_2}=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}} \tag{10-11}$$

两总体比例之差的区间估计

$$\bar{p}_1-\bar{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}+\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \tag{10-13}$$

当 $p_1=p_2=p$ 时，$\bar{p}_1-\bar{p}_2$ 的标准误差

$$\sigma_{\bar{p}_1-\bar{p}_2}=\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{10-14}$$

当 $p_1=p_2=p$ 时，p 的合并估计量

$$\bar{p}=\frac{n_1\bar{p}_1+n_2\bar{p}_2}{n_1+n_2} \qquad (10\text{-}15)$$

p_1-p_2的假设检验的检验统计量

$$z=\frac{\bar{p}_1-\bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad (10\text{-}16)$$

补充练习

38. Safegate Foods 公司正在重新设计其在全美各地超市的结账通道并提出了两种设计方案。在装有两种新系统的两家商店分别对顾客的结账时间进行了测试，得到了以下汇总数据。

系统 A	系统 B
$n_1=120$	$n_2=100$
$\bar{x}_1=4.1$ 分钟	$\bar{x}_2=3.4$ 分钟
$\sigma_1=2.2$ 分钟	$\sigma_2=1.5$ 分钟

在0.05的显著性水平下，检验两种系统的总体平均结账时间是否不同。哪一个系统更好？

40. 互助基金分为有佣金和无佣金两种类型。有佣金基金要求投资者以在基金中投资金额的一个百分比为依据，支付加盟费。无佣金基金则没有这个加盟费。一些金融顾问认为，因为有佣金的互助基金比无佣金的互助基金获得更高的平均收益率，因此有佣金的互助基金支付额外费用可能是值得的。分别抽取30个有佣金的互助基金和30个无佣金的互助基金各组成一个样本。采集的数据是5年期基金的年收益率。数据保存在本书所附光盘名为 Mutual 的数据集文件中。前5个有佣金的互助基金和前5个无佣金的互助基金的数据如下。

有佣金的互助基金	收益率
American national Growth	15.51
Arch Small Cap Equity	14.57
Bartlett Cap Basic	17.73
Calvert World International	10.31
Colonial Fund A	16.23
无佣金的互助基金	**收益率**
Amana Income Fund	13.24
Berger One Hundred	12.13
Columbia International Stock	12.17
Dodge & Cox Balanced	16.06
Evergreen Fund	17.61

a. 写出H_0和H_a，使得拒绝H_0会得出结论：5年期的有佣金互助基金有较高的平均年收益率。

b. 用数据集 Mutual 中的60个互助基金进行假设检验。p-值是多少？在$\alpha=0.05$的显著性水平下，你有何结论？

42. 2009年年初，经济正在经历一次衰退。但这次衰退对股票市场的影响如何呢？由15个公司组成的一个样本，样本中的每个公司在1月1日和4月30日的每股股票的价格如下所示（《华尔街日报》，2009年5月1日）。

公司	1月1日（美元）	4月30日（美元）
美国实用材料公司	10.13	12.21
纽约银行	28.33	25.48
雪佛兰	73.97	66.10
思科系统公司	16.30	19.32
可口可乐	45.27	43.05
康卡斯特	16.88	15.46
福特汽车	2.29	5.98
通用电气	16.20	12.65
强生	59.83	52.36
摩根大通	31.53	33.00
微软	19.44	20.26
甲骨文	17.73	19.34
辉瑞公司	17.71	13.36
菲利普莫里斯	43.51	36.18
宝洁	61.82	49.44

a. 四个月来每股股票的平均价格有什么变化？

b. 建立每股股票平均价格变化的90%的置信区间估计。解释这一结果。

c. 四个月来每股股票平均价格变化的百分比是多少？

d. 如果同样的百分比变化出现在随后的四个月以及再以后的四个月，2009年年底的每股股票平均价格将会是多少？

44. 一家大型汽车保险公司抽取了单身与已婚男性保险客户的样本，并记录了他们在过去的三年内保险索赔的次数。

单身保险客户	已婚保险客户
$n_1=400$	$n_2=900$
保险索赔次数 = 76	保险索赔次数 = 90

a. 在 $\alpha=0.05$ 的显著性水平下，检验单身与已婚保险客户的索赔率之间是否存在差异。

b. 建立两总体比例之差的 95% 的置信区间。

46. 预计在 2008 年 3 月期间，南卡罗莱纳州 Myrtle Beach 的公寓假期租用率将会上升（*The Sun News*，2008 年 2 月 29 日）。保存在本书所附光盘名为 Occupancy 文件中的数据来源于报纸，你可以复制使用。这些数据显示，在 2007 年 3 月和 2008 年 3 月的第一周，一个度假公寓随机样本的已租出的单元和未租出的单元的资料。

a. 估计在 2007 年 3 月和 2008 年 3 月的第一周，已租出单元的比例。

b. 建立比例之差的 95% 的置信区间。

c. 根据你的研究结果，与前一年同期相比，2008 年 3 月的租用率是否上升了？

案例 10-1　Par 公司

Par 公司是一家高尔夫球设备的主要制造商。管理人员认为引进某种耐磨损、寿命更长的高尔夫球会使 Par 公司的市场占有率增加，因此为了抗磨损，延长使用寿命，Par 的研究小组设计了一种带有涂层的新型高尔夫球。对涂层的测试已有获得成功的迹象。

一位研究者关注涂层对击球距离的影响。Par 希望新型耐磨的高尔夫球与目前使用的高尔夫球有相同的击球距离。为比较两种高尔夫球的击球距离，各取 40 只球来做距离测试。为了能将两种型号的高尔夫球平均距离的差异归因于两种球制作方法的不同，检验是用机械击球装置来完成的。检验结果如下，其中距离是按最接近的整数码测量。这些数据可供我们使用，它们被保存在本书的网站上。

型号		型号		型号		型号	
当前	新型	当前	新型	当前	新型	当前	新型
264	277	270	272	263	274	281	283
261	269	287	259	264	266	274	250
267	263	289	264	284	262	273	253
272	266	280	280	263	271	263	260
258	262	272	274	260	260	275	270
283	251	275	281	283	281	267	263
258	262	265	276	255	250	279	261
266	289	260	269	272	263	274	255
259	286	278	268	266	278	276	263
270	264	275	262	268	264	262	279

管理报告

1. 提出并介绍 Par 公司用于比较目前使用的和新型的高尔夫球击球距离的假设检验的基本原理。
2. 分析数据，得出假设检验的结论。检验的 p-值是多少？你对 Par 公司有何建议？
3. 对每种型号的数据给出描述性的统计汇总。
4. 每种型号的总体平均击球距离的 95% 置信区间是多少？两总体均值差的 95% 的置信区间是多少？
5. 你认为需要更大的样本容量和对高尔夫球做更多的检验吗？请讨论。

第11章 总体方差的统计推断

CHAPTER 11

实践中的统计

美国审计总署 ⊖
华盛顿特区

美国审计总署（GAO）是联邦政府立法机关的一个独立的、非官方的审计机构。GAO 的稽查员们确定当前的、拟议的联邦计划的有效性。为了履行他们的职责，稽查员们必须精通记录复审、立法调查以及统计分析技术。

在一个案例中，GAO 的稽查员们对一个净化全美河流湖泊的内务部计划进行研究。作为该计划的一部分，联邦政府的拨款要面向全美范围内的小城市。国会要求 GAO 确定该计划执行的效率。为此，GAO 检查了相关记录并走访了几家废物处理厂。

GAO 审计的目的之一是确认这些工厂的流出物（治理后的污水）达到一定的标准。另外，他们复查了样本的含氧量、pH 值和流出物中悬浮固体的数量。该计划的要求之一是在每个工厂进行日常的各种检验，并将收集的数据定期送往各州的工程部门。GAO 经过对数据的审查，得以确定流出物的各种特性是否在可接受的限度之内。

例如，仔细测定流出物的平均 pH 值。另外，需要复查上报的 pH 值的方差。关于流出物总体的 pH 值的方差进行如下的假设检验。

$$H_0:\sigma^2 = \sigma_0^2$$

$$H_a:\sigma^2 \neq \sigma_0^2$$

在本检验中，σ_0^2 表示一个运行良好的工厂其 pH 值所预期的总体方差。对某个特定的工厂，原假设被拒绝了，进一步的分析表明该厂 pH 值的方差显著地小于正常值。

稽查员们到这个特定的工厂检查测量设备并与该厂的管理者讨论有关统计数字。稽查员们发现因为操作员不知如何操作而并未使用测量设备。取而代之的是，某个工程师告诉操作员什么样的 pH 值是可接受的，然后操作员在没有做任何检验的情况下简单地记录了类似的数值。该厂异常低的方差导致拒绝 H_0。GAO 怀疑其他工厂也存在类似问题，便建议对操作员进行培训，以达到改进收集污染控制数据方式的目的。

在本章中，你将学会如何对一个和两个总体方差进行统计推断。我们将介绍两种新的分布：χ^2 分布和 F 分布，这两种分布可以用于对总体方差进行区间估计及假设检验。

在前面四章中，我们介绍了关于总体均值和总体比率的统计推断方法。在本章中我们把讨论扩展到总体方差的统计推断问题。关于方差能够为重要决策提供信息的案例，考虑灌装液体洗涤剂的生产过程。该生产过程的灌装机械经过调整，使得每个容器的平均灌装量为 16 盎司。尽管 16 盎司的平均值是所期望的，但是灌装量的方差也是很关键的。这就是说，即使灌装机械严格地调整到 16 盎司的平均值，我们也不能指望每个容器的灌装量恰好是 16 盎司。抽取一些容器组成样本，我们可以计算出容器中液体灌装量的样本方差，这个数值可以作为灌装生产过程中容器灌装量的总体方差的一个估计值。如果样本方差适度，生产过程可以继续。然而，如果样本方差过大，即使均值恰好是 16 盎司，也可能会发生过度灌装或灌装不足的情况，在这种情形下，灌装机械将被调整以减少容器灌装量的方差 ⊜。

在第 11.1 节我们考虑一个总体方差的统计推断。随后，我们将讨论对两个总体的方差进行推断的步骤。

11.1　一个总体方差的统计推断

样本方差

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{11-1}$$

是总体方差 σ^2 的点估计。在用样本方差作为推断总体方差的基础时，$(n-1)s^2/\sigma^2$ 的抽样分布起到了很大作

⊖ 作者感谢美国审计总署的 Art Foreman 先生和 Dale Ledman 先生，他们为“实践中的统计”提供了本案例。
⊜ 在许多制造业的应用中，控制过程的方差对保证产品质量是非常重要的。

用，这个抽样分布描述如下。

$(n-1)s^2/\sigma^2$ 的抽样分布

从正态总体中任意抽取一个容量为 n 的简单随机样本，则

$$\frac{(n-1)s^2}{\sigma^2} \tag{11-2}$$

的抽样分布服从自由度为 $n-1$ 的 χ^2 分布[⊖]。

图 11-1 显示了 $(n-1)s^2/\sigma^2$ 的抽样分布的几种可能的形式。

因为只要容量为 n 的简单随机样本取自正态总体，则 $(n-1)s^2/\sigma^2$ 的抽样分布就服从 χ^2 分布，我们可以利用 χ^2 分布对一个总体的方差建立区间估计和进行假设检验。

11.1.1 区间估计

为了阐述如何利用 χ^2 分布建立总体方差 σ^2 的置信区间估计，假定我们试图估计本章开始时提到的产品灌装过程的总体方差。抽取了 20 个容器组成一个样本，得到灌装量的样本方差 $s^2=0.0025$。但是，我们知道，由 20 个容器组成的样本无法提供生产过程中容器灌装量的总体方差的精确值。因此，我们考虑对总体方差进行区间估计。

我们用符号 χ^2_α 表示 χ^2 分布右侧面积或概率为 α 的数值。例如，对图 11-2 所示的自由度为 19 的 χ^2 分布，$\chi^2_{0.025}=32.852$ 表明有 2.5% 的 χ^2 分布值落在 32.852 右侧；$\chi^2_{0.975}=8.907$ 表明有 97.5% 的 χ^2 分布值落在 8.907 的右侧。对于 χ^2 分布的面积表或概率表很容易使用。参见表 11-1，可知自由度为 19 的 χ^2 分布的这些值是正确的（表中第 19 行）。附录 B 中的表 3 提供了更多的 χ^2 分布值。

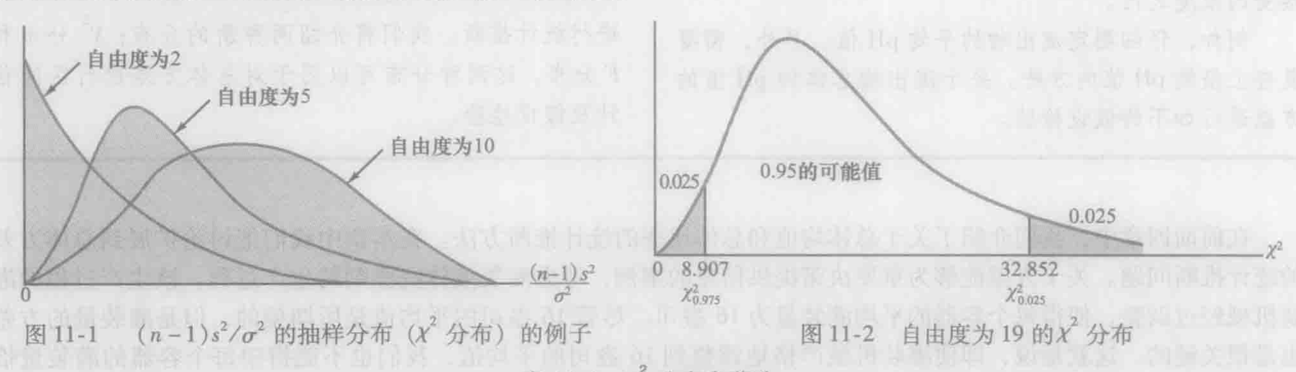


图 11-1 $(n-1)s^2/\sigma^2$ 的抽样分布（χ^2 分布）的例子

图 11-2 自由度为 19 的 χ^2 分布

表 11-1 χ^2 分布表节选

自由度	上侧面积							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345

⊖ χ^2 分布以取自正态总体的抽样为基础。

（续）

自由度	上侧面积							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.923
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.298	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

注：更完整的 χ^2 分布表见附录 B 中表 B-3。

从图 11-2 中可以看出，有 0.95 或 95% 的 χ^2 分布值位于 $\chi^2_{0.975}$ 与 $\chi^2_{0.025}$ 之间。这就是说，满足条件 $\chi^2_{0.975} \leqslant \chi^2 \leqslant \chi^2_{0.025}$ 的 χ^2 分布值的概率为 0.95。

正如式（11-2）所述，$(n-1)s^2/\sigma^2$ 服从 χ^2 分布，因此我们可以用 $(n-1)s^2/\sigma^2$ 代替 χ^2，得到

$$\chi^2_{0.975} \leqslant \frac{(n-1)s^2}{\sigma^2} \leqslant \chi^2_{0.025} \tag{11-3}$$

实际上，由于 $(n-1)s^2/\sigma^2$ 的所有可能取值落在区间 $\chi^2_{0.975}$ 到 $\chi^2_{0.025}$ 的概率为 0.95 或 95%，于是根据式（11-3）可以得到区间估计。现在我们需要对式（11-3）做一些代数运算以获得总体方差 σ^2 的区间估计。由式（11-3）左面的不等式，我们得到

$$\chi^2_{0.975} \leqslant \frac{(n-1)s^2}{\sigma^2}$$

即

$$\sigma^2\chi^2_{0.975} \leqslant (n-1)s^2$$

或

$$\sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{0.975}} \tag{11-4}$$

对式（11-3）右面的不等式做类似处理，得到

$$\frac{(n-1)s^2}{\chi^2_{0.025}} \leqslant \sigma^2 \tag{11-5}$$

综合式（11-4）和式（11-5），我们得到

$$\frac{(n-1)s^2}{\chi^2_{0.025}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{0.975}} \tag{11-6}$$

因为式（11-3）对于95%的 $(n-1)s^2/\sigma^2$ 的值成立，所以，式（11-6）给出了总体方差 σ^2 的95%的置信区间估计。

让我们回到求灌装量的总体方差的区间估计问题上来。我们记得由20个容器组成的样本得到的样本方差为 $s^2=0.002\,5$。由于样本容量为20，所以自由度为19。如图11-2所示，我们已经知道 $\chi^2_{0.975}=8.907$，$\chi^2_{0.025}=32.852$，将这些数值代入式（11-6），可以得到总体方差的区间估计如下。

$$\frac{19\times0.002\,5}{32.852} \leqslant \sigma^2 \leqslant \frac{19\times0.002\,5}{8.907}$$

或

$$0.001\,4 \leqslant \sigma^2 \leqslant 0.005\,3$$

对上式开平方根，我们得到总体标准差的95%的置信区间如下。[⊖]

$$0.038\,0 \leqslant \sigma \leqslant 0.073\,0$$

以上我们说明了利用 χ^2 分布建立总体方差与总体标准差的区间估计的过程。应特别注意，由于使用 $\chi^2_{0.975}$ 和 $\chi^2_{0.025}$，因此区间估计的置信系数为0.95。将式（11-6）推广到任意置信系数的一般情形，我们得到一个总体方差的区间估计如下。

一个总体方差的区间估计

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}} \tag{11-7}$$

式中，χ^2 值为基于自由度为 $n-1$ 的 χ^2 分布；$1-\alpha$ 为置信系数。

11.1.2 假设检验

用 σ_0^2 表示总体方差的假设值，关于总体方差有如下三种形式的假设检验：

$$\begin{array}{lll} H_0:\sigma^2 \geqslant \sigma_0^2 & H_0:\sigma^2 \leqslant \sigma_0^2 & H_0:\sigma^2 = \sigma_0^2 \\ H_a:\sigma^2 < \sigma_0^2 & H_a:\sigma^2 > \sigma_0^2 & H_a:\sigma^2 \neq \sigma_0^2 \end{array}$$

这三种形式与第9章和第10章中我们用于总体均值和比率的单侧和双侧假设检验的三种形式类似。

进行总体方差假设检验的步骤，是用总体方差的假设值 σ_0^2 和样本方差 s^2 来计算检验统计量 χ^2 的值。假设总体服从正态分布，则检验的统计量如下。

一个总体方差假设检验的检验统计量

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{11-8}$$

式中，χ^2 服从自由度为 $n-1$ 的 χ^2 分布。

⊖ 对总体方差置信区间的上限和下限开平方根，就能得到关于总体标准差的置信区间估计。

在计算出检验统计量χ^2 的数值之后，可使用 p-值法或者临界值法来确定原假设是否被拒绝。

让我们考虑如下例子，圣路易斯城市汽车公司鼓励其员工遵守时间，以在公众面前树立值得信赖的形象。作为一个规范制度，该公司要求各辆汽车的到站时间变化不大。就到站时间的方差而言，公司规定的标准是到站时间（以分钟计）的方差不超过 4。利用如下的假设检验，可以帮助公司确定到站时间的总体方差是否过大。

$$H_0: \sigma^2 \leqslant 4$$
$$H_a: \sigma^2 > 4$$

如果尝试性地假设 H_0 为真，即我们假定到站时间的总体方差满足公司规定的标准。如果当样本证据表明总体方差超过规定的标准时，则我们拒绝 H_0。从这一点上说，H_0 被拒绝意味着必须采取进一步的措施以减小总体方差。我们使用 $\alpha = 0.05$ 的显著性水平下进行假设检验。

假定在某个市中心车站随机抽取了 24 辆公共汽车的到站时间组成一个样本，得到样本方差为 $s^2 = 4.9$。假设到站时间的总体分布近似服从正态分布，则检验统计量的数值如下。

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(24-1) \times 4.9}{4} = 28.18$$

自由度为 $n-1=24-1=23$ 的χ^2 分布如图 11-3 所示。因为这是上侧检验，所以检验统计量$\chi^2 = 28.18$ 右侧的曲线下的面积就是检验的 p-值。

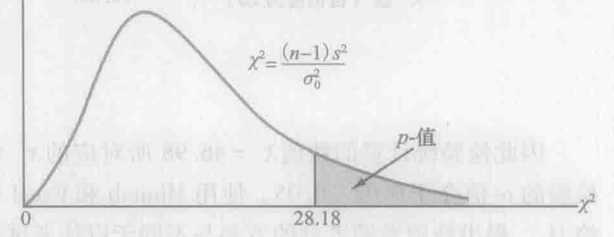


图 11-3　圣路易斯城市汽车公司例子的χ^2 分布

同 t 分布表类似，χ^2 分布表同样没有包含足够的细节使我们能够得到精确的 p-值。但是，我们可以利用χ^2 分布表得到 p-值的一个范围。例如，利用表 11-1，我们发现自由度为 23 的χ^2 分布有如下信息。

上侧面积	0.10	0.05	0.025	0.01
χ^2 值（自由度为 23）	32.007	35.172	38.076	41.638

↑
$\chi^2 = 28.18$

因为$\chi^2 = 28.18 < 32.007$，所以上侧面积（p-值）大于 0.10。由于 p-值 $> \alpha = 0.05$，所以我们不能拒绝原假设，即样本不支持到站时间的总体方差变大这个结论。

由于直接从χ^2 分布表很难得到精确的 p-值，但是诸如 Minitab 或 Excel 等计算机软件包可以帮助我们。本书后的附录 F 中描述了如何计算 p-值。在该附录中，我们得到$\chi^2 = 28.18$ 对应的 p-值为 0.209 1。

同其他假设检验程序一样，也可以使用临界值法得出假设检验的结论。当 $\alpha = 0.05$ 时，$\chi^2_{0.05}$ 给出了上侧检验的临界值。利用表 11-1，对于自由度为 23 的χ^2 分布，得到$\chi^2_{0.05} = 35.172$。因此，对到站时间例子的拒绝法则如下：

如果$\chi^2 \geqslant 35.172$，则拒绝 H_0

因为检验统计量的数值$\chi^2 = 28.18$，所以我们不拒绝原假设。

实践中，上侧检验最常用于关于总体方差的检验。在涉及到达时间、生产时间、灌装重量、零件尺寸等情形，一般都要求较小的方差，而较大的方差则不可能接受。在给出总体方差的最大容许值的情况下，我们可以检验原假设：总体方差小于或等于最大容许值；对立假设：总体方差大于最大容许值。根据这种检验的结构，一旦拒绝原假设，则显示总体方差过大，应该采取纠正措施。

同我们对总体均值和总体比率的检验一样，也可以进行其他形式的假设检验。让我们通过考察某机动车辆管理处所面临的问题，来说明关于总体方差的双侧检验。历史上，申请驾驶执照的个人考试分数的方差为 $\sigma^2 = 100$。现在考试采用了新型考题。该管理处的负责人希望新型考题的考分方差保持在原有水平上。为评价新型考题考分的方差，提出下面的双侧假设检验。

$$H_0: \sigma^2 = 100$$
$$H_a: \sigma^2 \neq 100$$

拒绝 H_0 表明方差发生了改变，因此为了使新型考题考分的方差保持原有水平，必须修订新型考题中的某些考题。一个由 30 名驾驶执照的申请者组成的样本将接受这种新型考题的考试。我们在 $\alpha=0.05$ 的显著性水平下进行假设检验。

由 30 份新型考题的考分组成的样本方差为 $s^2=162$，检验统计量 χ^2 的数值如下：

$$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}=\frac{(30-1)\times 162}{100}=46.98$$

现在让我们来计算 p-值。利用表 11-1 和自由度为 $n-1=29$，我们得到如下信息：

上侧面积	0.10	0.05	0.025	0.01
χ^2 值（自由度为 29）	39.087	42.557	45.722	49.588

$\chi^2=46.98$

因此检验统计量的数值 $\chi^2=46.98$ 所对应的 χ^2 分布上侧面积介于 0.01 ~ 0.025。将这些数值加倍，得到双侧检验的 p-值介于 0.02 ~ 0.05。使用 Minitab 和 Excel 可以得到精确的 p-值为 0.037 4。由于 p-值 $\leqslant\alpha=0.05$，我们拒绝 H_0，得出新型考题考分的方差与不同于以往考试分数的方差 $\sigma^2=100$。表 11-2 中给出了一个总体方差假设检验步骤的汇总。

表 11-2　一个总体方差假设检验的汇总

	下侧检验	上侧检验	双侧检验
假设	$H_0:\sigma^2\geqslant\sigma_0^2$ $H_a:\sigma^2<\sigma_0^2$	$H_0:\sigma^2\leqslant\sigma_0^2$ $H_a:\sigma^2>\sigma_0^2$	$H_0:\sigma^2=\sigma_0^2$ $H_a:\sigma^2\neq\sigma_0^2$
检验统计量	$\chi^2=\frac{(n-1)\ s^2}{\sigma_0^2}$	$\chi^2=\frac{(n-1)\ s^2}{\sigma_0^2}$	$\chi^2=\frac{(n-1)\ s^2}{\sigma_0^2}$
拒绝法则： p-值法	如果 p-值 $\leqslant\alpha$，则拒绝 H_0	如果 p-值 $\leqslant\alpha$，则拒绝 H_0	如果 p-值 $\leqslant\alpha$，则拒绝 H_0
拒绝法则： 临界值法	如果 $\chi^2\leqslant\chi^2_{(1-\alpha)}$，则拒绝 H_0	如果 $\chi^2\geqslant\chi^2_{\alpha}$，则拒绝 H_0	如果 $\chi^2\leqslant\chi^2_{(1-\alpha/2)}$ 或 $\chi^2\geqslant\chi^2_{\alpha/2}$，则拒绝 H_0

练　习

方法

☆2. 由 20 个项目组成一个样本，样本标准差为 5。

a. 计算总体方差的 90% 的置信区间。

b. 计算总体方差的 95% 的置信区间。

c. 计算总体标准差的 95% 的置信区间。

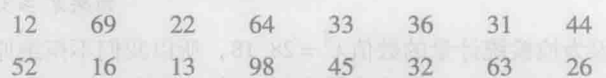
应用

4. 药品重量的方差在制药工业中非常关键。对某种特定药物（单位：克），由 18 份该药物组成的一个样本，得出样本方差 $s^2=0.36$。

a. 构造该药物重量的总体方差的 90% 置信区间。

b. 构造总体标准差的 90% 置信区间估计。

6. 美国人在万圣节服装和服饰上花费了近 70 亿美元（《华尔街日报》，2011 年 10 月 27 日）。16 个成年人组成一个样本，他们在万圣节服装上的花费总额数据如下（单位：美元）。



12	69	22	64	33	36	31	44
52	16	13	98	45	32	63	26

a. 成年人在万圣节服装上花费的总体均值的估计值是多少？

b. 样本标准差是多少？

c. 给出成年人在万圣节服装上花费总额的总体标准差的 95% 的置信区间估计。

8. 2009 年 3 月 4 日是 2009 年年初股票市场为数不多的几个好日子之一，该天的道琼斯工业股票平均价格指数大涨 149.82 点（《华尔街日报》，2009 年 3 月 25 日）。下表是 12 家公司该天股票价格变动的样本数据：

价格变动		价格变动	
公司	美元	公司	美元
Aflac	0.81	John. & John.	1.46
Bank of Am.	-0.05	Loews Cp	0.92
Cablevision	0.41	Nokia	0.21
Diageo	1.32	SmpraEngy	0.97
Flour Cp	2.23	Sunoco	0.52
Goodrich	0.30	Tyson Food	0.12

a. 计算该天价格变动的样本方差。

b. 计算该天价格变动的样本标准差。

c. 构建总体方差和总体标准差的95%的置信区间估计。

10.《消费者报告》用百分制的消费者满意度得分来评价全国的主要连锁店。假设根据过去的经验，消费者满意度得分的总体标准差为$\sigma=12$。2012年，在40个州拥有432家商铺的Costco是总体质量得到优秀的唯一一家连锁店。下面是15个Costco消费者满意度得分的样本数据。

95	90	83	75	95
98	80	83	82	93
86	80	94	64	62

a. Costco消费者满意度得分的样本均值是多少？

b. 样本方差是多少？

c. 样本标准差是多少？

d. 构造一个可以判定总体标准差$\sigma=12$是否被Costco拒绝的假设检验。在显著性水平为0.05下，你的结论如何？

12.《财富》杂志做了一项研究，发现其订阅者中拥有或租用车辆数量的方差为0.94。假定12名另一份杂志订阅者拥有或租用车辆的数据如下所示：2，1，2，0，3，2，2，1，2，1，0，1。

a. 计算这12名订阅者拥有或租用车辆数量的样本方差。

b. 对假设$H_0:\sigma^2=0.94$进行检验，以确定两份杂志订阅者拥有或租用车辆数的方差是否相同。在0.05的显著性水平下，你的结论如何？

11.2 两个总体方差的统计推断

在一些统计应用中，我们可能想比较两个不同生产工序生产出来的产品质量的方差，两种不同装配方法所需装配时间的方差，或者两种加热装置温度的方差。在比较两个总体方差时，我们将使用两个独立随机样本收集的数据，其中一个来自总体1，而另一个来自总体2，两个样本方差s_1^2和s_2^2是推断总体方差σ_1^2和σ_2^2的基础。当两个正态总体的方差相等（$\sigma_1^2=\sigma_2^2$）时，样本方差之比s_1^2/s_2^2有如下的抽样分布。

当$\sigma_1^2=\sigma_2^2$时，s_1^2/s_2^2的抽样分布

从两个方差相等的正态总体中分别抽取容量为n_1和n_2的两个独立的简单随机样本，则

$$\frac{s_1^2}{s_2^2} \tag{11-9}$$

的抽样分布服从分子自由度为n_1-1和分母自由度为n_2-1的F分布。s_1^2为取自总体1的容量为n_1的随机样本的样本方差，s_2^2为取自总体2的容量为n_2的随机样本的样本方差。㊀

图11-4是分子和分母自由度均为20的F分布曲线图形。从这张图上我们能看出，F分布是不对称的，而且F值永远不取负数。任何F分布的形状依赖于分子和分母的自由度。

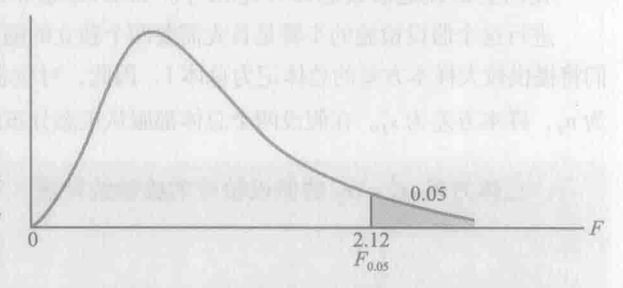


图11-4 分子与分母自由度都为20的F分布

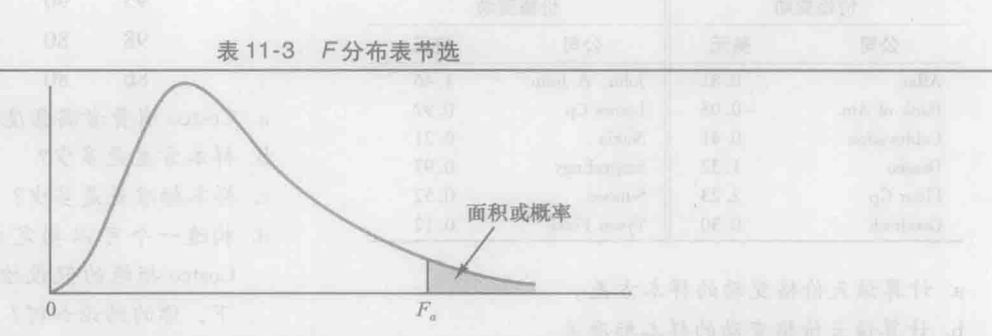
我们将用F_α表示F分布的上侧面积或概率为α时的值。例如，如图11-4所示，$F_{0.05}$表示对于分子和分母自由度都为20的F分布，其上侧面积为0.05。$F_{0.05}$这个特定值可以从F分布表中查出，表11-3是其中的一部分。根据分子和分母自由度都为20，以及相应的上侧面积为0.05，我们查出$F_{0.05}=2.12$。注意，对上侧面积为0.10，0.05，0.025和0.01，从该表中可以查出对应的F值。更为详细的F分布表见

㊀ F分布以取自两个正态总体的抽样为基础。



附录 B 的表4。

表 11-3 F分布表节选



分母自由度	上侧面积	分子自由度 10	15	20	25	30
10	0.10	2.32	2.24	2.20	2.17	2.16
	0.05	2.98	2.85	2.77	2.73	2.70
	0.025	3.72	3.52	3.42	3.35	3.31
	0.01	4.85	4.56	4.41	4.31	4.25
15	0.10	2.06	1.97	1.92	1.89	1.87
	0.05	2.54	2.40	2.33	2.28	2.25
	0.025	3.06	2.86	2.76	2.69	2.64
	0.01	3.80	3.52	3.37	3.28	3.21
20	0.10	1.94	1.84	1.79	1.76	1.74
	0.05	2.35	2.20	2.12	2.07	2.04
	0.025	2.77	2.57	2.46	2.40	2.35
	0.01	3.37	3.09	2.94	2.84	2.78
25	0.10	1.87	1.77	1.72	1.68	1.66
	0.05	2.24	2.09	2.01	1.96	1.92
	0.025	2.61	2.41	2.30	2.23	2.18
	0.01	3.13	2.85	2.70	2.60	2.54
30	0.10	1.82	1.72	1.67	1.63	1.61
	0.05	2.16	2.01	1.93	1.88	1.84
	0.025	2.51	2.31	2.20	2.12	2.07
	0.01	2.98	2.70	2.55	2.45	2.39

注：更完整的 F 分布值见附录 B 的表 B-4。

下面我们介绍如何用 F 分布进行两个总体方差的假设检验。我们以检验两个总体方差相等开始，所做的假设如下。

$$H_0:\sigma_1^2 = \sigma_2^2$$
$$H_a:\sigma_1^2 \neq \sigma_2^2$$

我们尝试性地假设总体方差相等。如果 H_0 被拒绝，我们将得出两个总体方差不相等的结论。

进行这个假设检验的步骤是首先需要两个独立的随机样本，且一个样本来自一个总体，然后计算两个样本方差。我们将提供较大样本方差的总体记为总体1，因此，对应总体1的样本容量为 n_1，样本方差为 s_1^2；对于总体2的样本容量为 n_2，样本方差为 s_2^2。在假设两个总体都服从正态分布的基础上，样本方差之比可以得到如下的 F 检验统计量。

总体方差 $\sigma_1^2=\sigma_2^2$ 的假设检验的检验统计量

$$F=\frac{s_1^2}{s_2^2} \tag{11-10}$$

将样本方差较大的总体记为总体1，则检验统计量服从分子自由度为 n_1-1，分母自由度为 n_2-1 的 F 分布。

因为进行检验的统计量 F 的分子是较大的样本方差 s_1^2，因此检验统计量的数值将出现在 F 分布的上侧，所以表 11-3 和附录 B 中的表4 只需要给出上侧面积或概率。如果我们不用这种方式构造检验统计量，可能需要用到下

侧面积或概率，在这种情形下，将需要额外的计算或更完整的 F 分布表。现在让我们考虑一个关于两个总体方差相等的假设检验例子。

Dullus 县立学校想更新明年校车服务的合同，必须从 Milbank 公司和 Gulf Park 公司这两个公共汽车公司中选择一个。我们将用到达时间或运送时间的方差作为衡量公共汽车公司服务质量的基本标准。较低的方差说明服务质量比较稳定而且水平比较高。如果两个公司的汽车到达时间的方差相等，Dullus 学校的管理者就会选择能提供较低价格的那个公司。然而，如果两个公司汽车到达时间的样本数据表明两个方差明显不等，管理者将优先考虑服务质量更好或方差更小的那个公司。对应的假设如下。

$$H_0:\sigma_1^2 = \sigma_2^2$$

$$H_a:\sigma_1^2 \neq \sigma_2^2$$

如果 H_0 被拒绝，则说明两个公司的服务质量不相同。我们在显著性水平 $\alpha=0.10$ 下进行假设检验。

选择 Milbank 公司的 26 次到达时间组成一个样本，得到样本方差为 $s_1^2=48$；选择 Gulf Park 公司的 16 次到达时间组成另一个样本，得到样本方差为 $s_2^2=20$。由于 Milbank 的样本方差较大，我们将 Milbank 视为总体 1。利用式（11-10），检验统计量的数值为

$$F=\frac{s_1^2}{s_2^2}=\frac{48}{20}=2.40$$

相应的 F 分布的分子自由度为 $n_1-1=26-1=25$，分母自由度为 $n_2-1=16-1=15$。

同其他假设检验程序一样，我们可以使用 p-值法或临界值法得出假设检验的结论。从表 11-3 中得到对分子自由度为 25，分母自由度为 15 的 F 分布上侧面积所对应的 F 值如下。

上侧面积	0.10	0.05	0.025	0.01
F 值（$df_1=25$，$df_2=15$）	1.89	2.28	2.69	3.28

↑ $F=2.40$

因为 $F=2.40$ 介于 2.28 ~ 2.69，因此 F 分布的上侧面积就介于 0.025 ~ 0.05。既然这是双侧检验，我们将上侧面积加倍，得到 p-值介于 0.05 ~ 0.10。对于这个检验，我们选择 $\alpha=0.10$ 作为显著性水平，由于 p-值 $<\alpha=0.10$，因此原假设被拒绝，这个结果导致我们得出两个公共汽车服务运送时间的方差是不同的结论。因此 Dullus 学校的管理者就可以优先考虑服务更好或方差较小的 Gulf Park 公司。

我们利用 Excel 或 Minitab 可以得到检验统计量 $F=2.40$ 对应的双侧 p-值为 0.081 1。由于 $0.0811<\alpha=0.10$，所以两个总体方差相等的原假设被拒绝。

在显著性水平 $\alpha=0.10$ 下，用临界值法进行双侧假设检验。我们使用分布每一侧的面积 $\alpha/2=0.10/2=0.05$ 来选择临界值。因为利用式（11-10）计算的检验统计量的数值永远在分布的上侧，所以我们只需要确定上侧临界值。从表 11-3 中我们看到 $F_{0.05}=2.28$。因此，尽管我们使用的是双侧检验，拒绝法则可陈述如下。

如果 $F\geqslant 2.28$，则拒绝 H_0

因为 $F=2.40>2.28$，所以我们拒绝 H_0，并得出两个公共汽车服务运送时间的方差不相等的结论。

涉及两个总体方差的单侧检验也是可能做到的，在这种情形下，我们用 F 分布来确定一个总体的方差是否显著大于另一个总体的方差。关于两个总体方差的单侧假设检验可永远写成上侧检验的形式：[⊖]

$$H_0:\sigma_1^2 \leqslant \sigma_2^2$$

$$H_a:\sigma_1^2 > \sigma_2^2$$

假设检验的这种形式永远讨论 F 分布的上侧 p-值和临界值。因此，只需要上侧 F 值，简化了计算和查 F 分布表。

让我们通过考虑民意调查来说明使用 F 分布进行两个总体方差的单侧检验。由 31 名男性和 41 名女性组成的

⊖ 两个总体方差的单侧假设检验可以永远写成上侧检验的形式，这种方法消除了对下侧 F 值的需要。

样本将用于研究他们对当前政治问题的态度。研究者想通过样本数据检验，女性对政治问题的态度的方差是否比男性要大。在前边给出的单侧假设检验的形式中，用总体1代表女性，用总体2代表男性，假设检验将表示如下。

$$H_0: \sigma^2_{\text{women}} \leqslant \sigma^2_{\text{men}}$$

$$H_a: \sigma^2_{\text{women}} > \sigma^2_{\text{men}}$$

如果 H_0 被拒绝，研究者就能得到他所需的统计支持，从而作出女性对政治问题的态度具有较大方差的结论。

由于用女性的样本方差作为分子而用男性的样本方差作为分母，所以 F 分布的分子自由度为 $n_1 - 1 = 41 - 1 = 40$，分母自由度为 $n_2 - 1 = 31 - 1 = 30$。我们使用显著性水平 $\alpha = 0.05$ 来进行假设检验。调查结果表明，女性的样本方差 $s_1^2 = 120$，男性的样本方差 $s_2^2 = 80$。则检验的统计量如下：

$$F = \frac{s_1^2}{s_2^2} = \frac{120}{80} = 1.50$$

参考附录B中表4，我们发现对于分子自由度为40，分母自由度为30的 F 分布有 $F_{0.10} = 1.57$。因为检验统计量 $F = 1.50 < 1.57$，上侧面积必然大于0.10。所以，我们可以得出 p-值大于0.10。利用Minitab或Excel得到 p-值为0.1256。因为 p-值 $> \alpha = 0.05$，不能拒绝 H_0，因此，样本结果不支持女性对政治问题态度的方差比男性大的结论。表11-4中给出了两个总体方差的假设检验的汇总。

表11-4 两个总体方差假设检验的汇总

	上侧检验	双侧检验
假设	$H_0: \sigma_1^2 \leqslant \sigma_2^2$ $H_a: \sigma_1^2 > \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_a: \sigma_1^2 \neq \sigma_2^2$ 注：总体1拥有较大的样本方差
检验统计量	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{s_1^2}{s_2^2}$
拒绝法则：p-值方法	如果 p-值 $\leqslant \alpha$，拒绝 H_0	如果 p-值 $\leqslant \alpha$，拒绝 H_0
拒绝法则：临界值法	如果 $F \geqslant F_\alpha$，则拒绝 H_0	如果 $F \geqslant F_{\alpha/2}$，则拒绝 H_0

注释和评论

研究人员证实了 F 分布对于正态总体的假定是敏感的。只有当两个总体至少近似正态分布的假定是合理的情况下，才能使用 F 分布。

练习

方法

14. 从总体1抽取16个项目组成一个样本，得到样本方差 $s_1^2 = 5.8$；从总体2抽取21个项目组成一个样本，得到样本方差 $s_2^2 = 2.4$。在显著性水平0.05下，检验下列假设：

$$H_0: \sigma_1^2 \leqslant \sigma_2^2$$

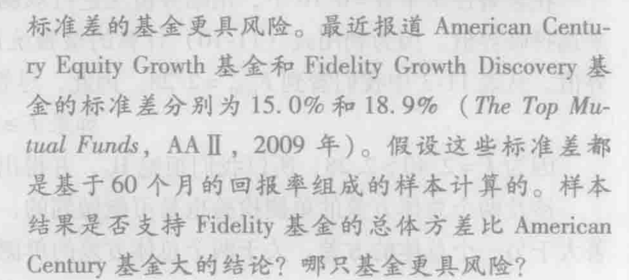
$$H_a: \sigma_1^2 > \sigma_2^2$$

a. 使用 p-值法，你有何结论？

b. 使用临界值法重复上述检验。

应用

16. 投资者通常用共同基金月回报率的标准差来度量基金的风险；有较大标准差的基金被认为比有较少标准差的基金更具风险。最近报道American Century Equity Growth基金和Fidelity Growth Discovery基金的标准差分别为15.0%和18.9%（*The Top Mutual Funds*，AAⅡ，2009年）。假设这些标准差都是基于60个月的回报率组成的样本计算的。样本结果是否支持Fidelity基金的总体方差比American Century基金大的结论？哪只基金更具风险？



18. *Barron's* 搜集了1000名顶级财政顾问的数据（*Barron's*，2009年2月9日），其中美林有239人，摩根士丹利有121人。一个由16名美林顾问和10名摩根士丹利顾问组成的样本表明，顾问们管理着许多在资金总量上有很大方差的大账户。美林顾问们管理资金总量的标准差为5.87亿美元，摩根士丹

利顾问管理资金总量的标准差为 4.89 亿美元。在 $\alpha=0.10$ 下进行一个假设检验，以确定两家公司管理的资金总量的总体方差是否有显著差异。对于两家公司在管理资金总量上的差异，你有何结论？

20. 根据 Romac 对薪金调查提供的资料，注册会计师事务所资深人员年薪的方差大约为 2.1，而注册会计师事务所管理人员年薪的方差大约为 11.1，薪金数据以 1 000 美元计。假设薪金数据是基于 25 名资深人员和 26 名管理人员组成的样本，检验假设：二者薪金的总体方差相等。在显著性水平 0.05 下，你有何结论？

22. 一项研究假设在湿滑路面上汽车刹车距离的方差明显比在干燥路面上汽车刹车距离的方差要大。在调查研究中，以同样速度行驶的 16 辆汽车分别在湿滑路面上和干燥路面上检测刹车距离。在湿滑路面上，刹车距离的标准差为 32 英尺；在干燥路面上，标准差为 16 英尺。
    a. 在显著性水平 0.05 下，样本数据是否能够证明湿滑路面上刹车距离的方差比干干燥路面上刹车距离方差大的结论？p-值是多少？
    b. 就驾驶安全性方面的建议而言，你的统计结论的含义是什么？

小　结

本章我们介绍了可以用于进行总体方差推断的统计方法。在这个过程中，我们介绍了两种新的概率分布：χ^2 分布和 F 分布。χ^2 分布可以作为一个正态总体方差的区间估计和假设检验的基础。

我们阐述了在两个总体方差的假设检验中 F 分布的运用。特别地，我们说明了从两个方差相等（$\sigma_1^2=\sigma_2^2$）的正态总体分别抽取容量为 n_1 和 n_2 的独立简单随机样本，其样本方差之比 s_1^2/s_2^2 的抽样分布服从分子自由度为 n_1-1 和分母自由度为 n_2-1 的 F 分布。

重要公式

一个总体方差的区间估计

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}} \qquad (11\text{-}7)$$

一个总体方差假设检验的检验统计量

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \qquad (11\text{-}8)$$

总体方差 $\sigma_1^2=\sigma_2^2$ 的假设检验的检验统计量

$$F = \frac{s_1^2}{s_2^2} \qquad (11\text{-}10)$$

补充练习

24. 平均而言，首次公开发行（IPO）股票的价格是折价的。标准差在折价 - 溢价指标中用于衡量其散布或变动情况。由 13 只在多伦多股票交易所交易的加拿大 IPO 组成一个样本，得到的样本标准差为 14.95。构建折价 - 溢价指标的总体标准差的 95% 的置信区间估计。

26. 部件变异性在滚珠轴承生产过程中至关重要。滚珠轴承尺寸的方差过大将导致滚动失败以及磨损过快。滚珠尺寸以英寸为单位，生产标准要求最大方差为 0.000 1。由 15 个滚珠组成一个样本，得到样本标准差为 0.014 英寸。
    a. 在 $\alpha=0.10$ 下，确定样本是否表明：最大容许方差已经被超出。
    b. 计算滚珠总体方差的 90% 的置信区间估计。

28. City Trucking 公司声称，对其日常客户运送货物的交货时间是始终如一的。一个由 22 次运送货物所组成的样本，得到样本方差为 1.5。在 $\alpha=0.10$ 下，检验以确定能否拒绝 $H_0:\sigma^2\leqslant 1$。

30. 乘坐某航空公司航班的乘客数量的样本标准差为 8 名乘客。总体标准差的 95% 的置信区间为 5.86 ~ 12.62 名乘客。
    a. 在该统计分析中，所用的样本容量是 10 还是 15？
    b. 假设一个由 25 个航班组成的样本得到样本标准差为 $s=8$ 名乘客，则总体标准差的置信区间将会有何变化？计算样本容量为 25 的 σ 的 95% 的置信区间估计。

32. 修完一门大学财务会计课程的 352 名学生 GPA 的标准差为 0.940。退选这门课程的 73 名学生 GPA 的标准差为 0.797。这些数据是否表明：修完与退选财务会计课程的学生 GPA 的方差之间存在显著

差异？取显著性水平为 0.05（注：自由度为 351 和 72 的 $F_{0.025}$ 数值为 1.466）。

34. 对两种新的装配方法进行检测，并报告装配时间的方差。在 $\alpha=0.10$ 下，检验两个总体方差是否相等。

	方法 A	方法 B
样本容量	$n_1=31$	$n_2=25$
样本方差	$s_1^2=25$	$s_2^2=12$

案例 空军训练计划

空军电子学引导性教程采用一种个人化教学系统，每位学生观看讲座录像，然后给以程式化的教材。每位学生独立学习直至完成训练并通过考试。人们关心的问题是学生完成训练计划的速度的不同。有些学生能够相当快地完成程式化教材，而另一些学生在教材上需要花费较长的时间，甚至需要加班加点才能完成课程。学得较快的学生必须等待学得较慢的学生完成引导性教程后才能一起进行其他方面的训练。

建议的替代系统是使用计算机辅助教学。在这种方法中，所有的学生观看同样的讲座录像，然后每位学生被指派到一个计算机终端来接受进一步的训练。在整个教程的自我训练过程中，由计算机指导学生独立操作。

为了比较建议的和当前的教学方法，刚入学的 122 名学生被随机地安排到这两种教学系统中。61 名学生使用当前程式化教材，而另外 61 名学生使用建议的计算机辅助方法。记录每位学生的学习时间（单位：小时）。下列数据存在光盘中名为 Training 的数据集中。

采用当前训练方法完成教程的时间（小时）										
76	76	77	74	76	74	74	77	72	78	73
78	75	80	79	72	69	79	72	70	70	81
76	78	72	82	72	73	71	70	77	78	73
79	82	65	77	79	73	76	81	69	75	75
77	79	76	78	76	76	73	77	84	74	74
69	79	66	70	74	72					
采用建议的计算机辅助方法完成教程的时间（小时）										
74	75	77	78	74	80	73	73	78	76	76
74	77	69	76	75	72	75	72	76	72	77
73	77	69	77	75	76	74	77	75	78	72
77	78	78	76	75	76	76	75	76	80	77
76	75	73	77	77	77	79	75	75	72	82
76	76	74	72	78	71					

管理报告

1. 利用适当的描述统计学方法汇总每种方法的训练时间数据。根据样本资料，你能观察到有何相似之处和差异？
2. 利用第 10 章的方法评价两种方法总体均值之间的差异，讨论你的结论。
3. 计算每一种训练方法的标准差与方差。进行两种训练方法总体方差相等的假设检验，讨论你的结论。
4. 关于两种方法之间的差异，你能得到什么结论？你有何建议？请作出解释。
5. 对于将来要使用哪种训练计划，在作出最终决定之前你是否还需要其他数据或者检验？

第 12 章

多个比率的比较、独立性及拟合优度检验

CHAPTER 12

实践中的统计

联合劝募协会[⊖]

纽约州，罗切斯特

大罗切斯特地区的联合劝募协会（United Way）是一个非营利性组织，该组织致力于满足社区最重要的人类关怀需求，从而提高它所服务的7个县全部居民的生活质量。

每年春季，联合劝募协会/红十字会举行募捐活动，所得资金用于超过200个服务机构所提供的几百个服务项目。这些服务机构提供能满足人类各种需求的服务，如身体的、精神的、社会的，并且为各种年龄、背景和经济条件的人提供服务。

为进一步了解社区的慈善观念，大罗切斯特地区的联合劝募协会决定进行一项调查。为了得到慈善观念的初步资料，他们走访了专业人员、服务人员和一般工人这些重点群体。根据这些资料，设计出调查问卷。问卷经过预调查、修订之后，分发给440个人。

从搜集到的数据得到各种描述性统计量，包括频数分布和交叉分组表。分析该数据的一个重要部分是独立性χ^2检验，这些统计检验的一个用途是确定行政开支观念是否与职业独立。

独立性检验的假设为：

H_0：联合劝募协会行政开支观念与调查对象的职业独立。

H_a：联合劝募协会行政开支观念与调查对象的职业不独立。

调查中，两个问题提供了用于统计检验的分类型数据。其中一个问题得到了有关行政开支占资金比例的数据（10%以下、11%～20%及21%以上），另一个问题是调查对象的职业。

独立性检验导致原假设被拒绝，因此可以断定联合劝募协会的行政开支观念的确与调查对象的职业有关。实际的行政开支小于9%，但是有35%的调查对象认为行政开支超过21%。因此，很多人有不准确的行政开支观念。在这一群体中，生产线工人、售货员和专业技术人员比群体中其他人有更不准确的行政开支观念。

社区观念的研究帮助罗切斯特地区的联合劝募协会调整项目和筹集资金活动。在本章你将学会如何进行如上所述的检验。

在第9、10和11章中，我们介绍了关于一个和两个总体的总体均值、比率和方差的假设检验的统计推断方法。本章，我们介绍另外3种假设检验方法，它们扩充了我们对总体进行统计推断的能力。

本章进行的假设检验所使用的检验统计量都基于χ^2分布。在所有的情形中，数据都是分类型的。这些χ^2检验是多用途的，并由检验下列的应用扩充了假设。

1. 检验三个或多个总体比率的相等性。
2. 检验两个分类变量的独立性。
3. 检验一个总体的概率分布是否服从一个特殊的历史或理论的概率分布。

我们从考虑三个或多个总体比率的相等性开始。

12.1 三个或多个总体比率的相等性的检验

在第10.2节，我们介绍了两个总体比率的统计推断方法，其假设检验的结论基于标准正态检验统计量z。这里我们展示如何使用检验统计量χ^2，对三个或多个总体比率的相等性进行统计推断。使用如下记号：

p_1——总体1的总体比率；

p_2——总体2的总体比率；

…

p_k——总体k的总体比率。

⊖ 作者感谢联合劝募协会的市场营销顾问Philip R. Tyler博士，他为“实践中的统计”提供了本案例。

关于 k 个总体比率相等性的假设陈述如下：

$$H_0: p_1 = p_2 = \cdots = p_k$$

H_a：所有总体比率不全相等

如果样本数据和 χ^2 检验计算结果表明 H_0 没有被拒绝，我们不能断定 k 个总体比率有差异。然而，如果样本数据和 χ^2 检验计算结果表明 H_0 被拒绝，我们有统计证据得出 k 个总体比率不全相等的结论；也就是说，一个或多个总体比率与其他总体比率不相等。可以进行进一步的分析，以得出哪个或哪些总体比率与其他总体比率有显著差异。我们由一个应用来说明 χ^2 检验。

像 J. D. Power and Associates 这样的组织，会以使用车主可能再次购买某款汽车的比率作为顾客汽车品牌忠诚度的指标。车主可能再次购买某款汽车的比率较大被认为有较大的顾客品牌忠诚度。假设在某次特殊的研究中，我们希望比较雪佛兰羚羊、福特 Fusion 和本田雅阁三种汽车的顾客品牌忠诚度。当前三种汽车的车主构成研究的三个总体，感兴趣的三个总体比率如下：

p_1——雪佛兰羚羊车主总体可能再次购买羚羊的比率；

p_2——福特 Fusion 车主总体可能再次购买 Fusion 的比率；

p_3——本田雅阁车主总体可能再次购买雅阁的比率。

假设陈述如下：

$$H_0: p_1 = p_2 = p_3$$

H_a：所有总体比率不全相等

为了进行这个假设检验，我们首先从三个总体中的每一个总体抽取车主样本，从而我们将得到一个雪佛兰羚羊车主的样本、一个福特 Fusion 车主的样本和一个本田雅阁车主的样本。每一个样本提供的分类型数据表明回答者是否可能或不可能再次购买该品牌汽车。125 名雪佛兰羚羊车主、200 名福特 Fusion 车主和 175 名本田雅阁车主的样本数据汇总结果如表 12-1 所示。[⊖] 表 12-1 中的两行对应的回答为“是”和“否”，3 列中的每一列对应于一个总体。汇总在表 12-1 中的 6 个单元格观察频数对应于可能再次购买的回答和三个总体的每一个组合。

表 12-1　三个汽车车主的总体可能再次购买的样本结果（观察频数）

		汽车车主			合计
		雪佛兰羚羊	福特 Fusion	本田雅阁车	
可能再次购买	是	69	120	123	312
	否	56	80	52	188
	合计	125	200	175	500

利用表 12-1，我们看到 125 名雪佛兰羚羊车主中，有 69 名表明他们可能再次购买雪佛兰羚羊。200 名福特 Fusion 车主中有 120 名以及 175 名本田雅阁车主中有 123 名表明他们可能再次购买他们当前品牌的汽车。同时，综合 3 个样本，本研究中的 500 名车主中有 312 名表明他们可能再次购买他们当前品牌的汽车。现在的问题是我们如何分析表 12-1 中的数据，以确定假设 $H_0: p_1 = p_2 = p_3$ 是否被拒绝？

表 12-1 中的数据是描述可能再次购买的回答与车主总体的 6 个组合的 6 个单元格中每一个的观察频数。如果在假设 H_0 为真时，我们能确定期望频数，我们就可以利用检验统计量 χ^2 来确定观察频数和期望频数之间是否存在显著差异。如果观察频数和期望频数之间存在显著差异，假设 H_0 将被拒绝，就有所有总体比率不全相等的证据。

表中 6 个单元格的期望频数基于以下基本原理。首先，我们假定总体比率相等的原假设为真，然后，我们注意到在全部 500 名车主的样本中，总共有 312 名车主表明他们可能再次购买他们当前品牌的汽车。因此，312/500 = 0.624 是可能再次购买他们当前品牌汽车的车主的总样本比率。如果 $H_0: p_1 = p_2 = p_3$ 为真，0.624 将是每

⊖ 在类似这样的研究中，对每一个总体我们常常使用相同的样本容量。在这个例子中，我们选择了不同的样本容量，以表明 χ^2 检验不局限于 k 个总体中的每一个的样本容量都相等。

一种汽车车主总体回答可能再次购买比率的最佳估计值。因此，如果 H_0 为真，我们将期望125名雪佛兰羚羊车主中的0.624，或 $0.624\times125=78$ 名车主表明他们可能再次购买羚羊。利用总样本比率0.624，我们期望200名福特Fusion车主中的 $0.624\times200=124.8$ 名以及175名本田雅阁车主中的 $0.624\times175=109.2$ 名，分别回答他们可能再次购买相应的汽车品牌。

将这个方法推广，让我们计算表中位于第 i 行和第 j 列的单元格的期望频数 e_{ij}，使用这个记号，现在重新考虑对可能再次购买的回答为“是”（第1行）的雪佛兰羚羊车主（第1列）的期望频数的计算，即期望频数 e_{11}。

注意回答为“是”的合计数为312（第1行合计数），雪佛兰羚羊车主的样本容量为的125（第1列合计数），总样本容量为500。遵循上一段的逻辑，我们可以得到

$$e_{11}=\left(\frac{\text{第1行合计数}}{\text{总样本容量}}\right)\times\text{第1列合计数}=\left(\frac{312}{500}\right)\times125=0.624\times125=78$$

由上面表达式的第一部分，我们得到

$$e_{11}=\frac{\text{第1行合计数}\times\text{第1列合计数}}{\text{总样本容量}}$$

推广之，可以得到在假设 H_0 为真时，用于计算期望频数的公式如下。

在假定 H_0 为真时的期望频数

$$e_{ij}=\frac{\text{第}i\text{行合计数}\times\text{第}j\text{列合计数}}{\text{总样本容量}}\tag{12-1}$$

利用式（12-1），我们看到对本田雅阁车主（第3列）回答为“是”（第1行）的期望频数 e_{13} =（第1行合计数）×（第3列合计数）/总样本容量 $=312\times175/500=109.2$。利用式（12-1），得到的期望频数如表12-2所示。

表12-2 如果 H_0 为真，三个汽车车主的总体可能再次购买的期望频数

		汽车车主			合计
		雪佛兰羚羊	福特 Fusion	本田雅阁车	
可能再次购买	是	78	124.8	109.2	312
	否	47	75.2	65.8	188
	合计	125	200	175	500

比较表12-1中的观察频数和表12-2中的期望频数的检验方法，涉及下列统计量 χ^2 的计算。

检验统计量 χ^2

$$\chi^2=\sum_i\sum_j\frac{(f_{ij}-e_{ij})^2}{e_{ij}}\tag{12-2}$$

式中，f_{ij} 为第 i 行和第 j 列的单元格的观察频数；e_{ij} 代表在假定 H_0 为真时第 i 行和第 j 列的单元格的期望频数。

注意：在涉及 k 个总体比率相等性的 χ^2 检验中，上述检验统计量服从自由度为 $k-1$ 的 χ^2 分布，其中每个单元格的期望频数都大于或等于5。

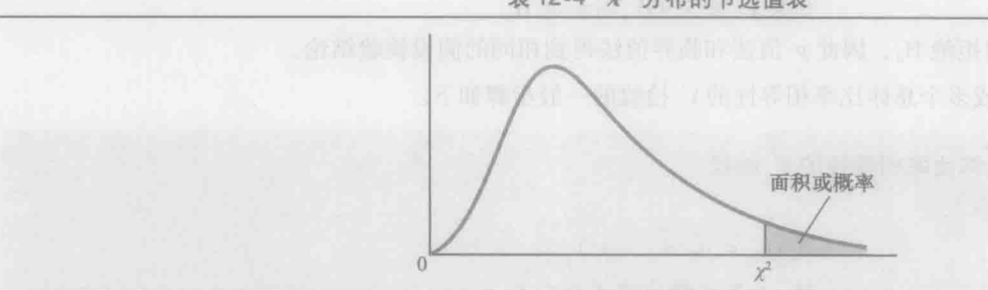
再次查看表12-2中的期望频数，我们发现表中每个单元格的期望频数都大于5。因此，我们进行检验统计量 χ^2 的计算。在表12-3中给出了计算检验统计量值的必要计算结果。这里，我们得到检验统计量的数值为 $\chi^2=7.89$。

表 12-3 检验总体比率相等性的检验统计量χ^2 的计算

是否可能再次购买	汽车车主	观察频数 (f_{ij})	期望频数 (e_{ij})	差 ($f_{ij}-e_{ij}$)	差的平方 ($f_{ij}-e_{ij})^2$	差的平方除以期望频数 $(f_{ij}-e_{ij})^2/e_{ij}$
是	羚羊	69	78.0	-9.0	81.00	1.04
是	Fusion	120	124.8	-4.8	23.04	0.18
是	雅阁	123	109.2	13.8	190.44	1.74
否	羚羊	56	47.0	9.0	81.00	1.72
否	Fusion	80	75.2	4.8	23.04	0.31
否	雅阁	52	65.8	-13.8	190.44	2.89
	合计	500	500			$\chi^2=7.89$

为了确定$\chi^2=7.89$ 是否导致我们拒绝 $H_0:p_1=p_2=p_3$，你需要了解并参考χ^2 分布值。表 12-4 显示了χ^2 分布的一般形状，但要注意一个特殊的χ^2 分布的形状依赖于其自由度的数值。表中显示了自由度 1～15 的χ^2 分布上侧面积分别为 0.10，0.05，0.025，0.01 和 0.005 的数值。这个版本的χ^2 分布表能够让你进行本章所介绍的假设检验。

表 12-4 χ^2 分布的节选值表



自由度	上侧面积 0.10	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801

由于表 12-2 的期望频数基于假定 $H_0:p_1=p_2=p_3$ 为真，观察频数 f_{ij} 与期望频数 e_{ij} 一致，在式（12-2）中将得到 $(f_{ij}-e_{ij})^2$ 较小的数值。如果这样，检验统计量χ^2 的数值将相对较小，且 H_0 不能被拒绝。另一方面，如果观察频数与期望频数之间的差异较大，$(f_{ij}-e_{ij})^2$ 的数值以及检验统计量的数值将较大，这时，总体比率相等的原假设可能被拒绝。因此，总体比率相等性的χ^2 检验永远是一个上侧检验[⊖]，当检验统计量位于χ^2 分布的上侧时，得

⊖ 本节介绍的χ^2 检验永远是一个单侧检验，H_0 的拒绝域在χ^2 分布的上侧。

到 H_0 的拒绝域。

我们可以用适当的χ^2 分布的上侧面积和 p-值法来确定原假设是否被拒绝。在汽车品牌忠诚度的研究中，三个车主总体表明适当的χ^2 分布的自由度为 $k-1=3-1=2$。利用χ^2 分布表的第 2 行，我们有如下结果：

上侧面积	0.10	0.05	0.025	0.01	0.005
χ^2 值（$df=2$）	4.605	5.991	7.378	9.210	10.597

$\chi^2=7.89$

我们看到$\chi^2=7.89$ 的上侧面积介于 0.025 ~ 0.01，因此，相对应的上侧面积或 p-值必然介于 0.025 ~ 0.01。由于 p-值≤0.05，我们拒绝 H_0，并得出三个总体比率不全相等，因此在雪佛兰羚羊、福特 Fusion 和本田雅阁车主中品牌忠诚度之间存在差异。利用附录 F 中提供的 Minitab 或 Excel 程序，对于自由度为 2，$\chi^2=7.89$ 对应的 p-值 = 0.019 3。

不使用 p-值法，我们可以使用临界值法得到相同的结论。对于 $\alpha=0.05$ 和自由度为 2，检验统计量χ^2 的临界值为$\chi^2=5.991$。上侧检验的拒绝法则变为：

$$\text{如果}\ \chi^2 \geqslant 5.991\text{，则拒绝}\ H_0$$

由于 7.89≥5.991，我们拒绝 H_0，因此 p-值法和临界值法得到相同的假设检验结论。

我们总结用于进行三个或多个总体比率相等性的χ^2 检验的一般步骤如下。

对于 $k\geqslant 3$ 个总体，总体比率相等性的χ^2 检验

1. 建立原假设与备选假设。

$$H_0: p_1 = p_2 = \cdots = p_k$$

$$H_a\text{：所有总体比率不全相等}$$

2. 从每一个总体中选择一个随机样本，并记录 2 行 k 列的表中的观察频数 f_{ij}。
3. 假定原假设为真，计算期望频数 e_{ij}。
4. 如果每一个单元格的期望频数 e_{ij} 都大于或等于 5，计算检验统计量：

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij}-e_{ij})^2}{e_{ij}}$$

5. 拒绝法则：

p-值法：如果 p-值 $\leqslant \alpha$，则拒绝 H_0

临界值法：如果 $\chi^2 \geqslant \chi^2_\alpha$，则拒绝 H_0

式中，χ^2 分布的自由度为 $k-1$，α 为检验的显著性水平。

多重比较方法

我们使用χ^2 检验得到三个汽车车主总体的总体比率不全相等。因此，有些总体比率之间存在差异，而且研究表明雪佛兰羚羊、福特 Fusion 和本田雅阁车主中顾客品牌忠诚度不全相同。为了说明总体比率之间存在的差异，我们先计算三个样本比率如下：

品牌忠诚度的样本比率

雪佛兰羚羊 $\bar{p}_1 = 69/125 = 0.552\,0$

福特 Fusion $\bar{p}_2 = 120/200 = 0.600\,0$

本田雅阁 $\bar{p}_3 = 123/175 = 0.702\,9$

由于χ^2检验表明总体比率不全相等，因此，我们进行尝试性确定哪些总体比率之间存在差异是合理的。为此，我们依靠用于对所有成对的总体比率进行统计检验的多重比较方法。接下来，我们讨论一种多重比较方法——著名的**Marascuilo方法**（Marascuilo procedure）。对所有成对的总体比率，这是一种相对简单的方法。我们将用汽车顾客品牌忠诚度的研究来说明这种多重比较检验方法所需要的计算。

首先，我们计算研究中每一对总体的样本比率之差的绝对值。在三个汽车品牌忠诚度研究的总体中，我们比较总体1和2、总体1和3以及总体2和3的样本比率如下：

雪佛兰羚羊与福特Fusion

$$|\bar{p}_1-\bar{p}_2| = |0.5520-0.6000| = 0.0480$$

雪佛兰羚羊与本田雅阁

$$|\bar{p}_1-\bar{p}_3| = |0.5520-0.7029| = 0.1509$$

福特Fusion与本田雅阁

$$|\bar{p}_2-\bar{p}_3| = |0.6000-0.7029| = 0.1029$$

第2步，我们选择显著性水平，并用下列表达式计算每一对比较值相对应的临界值。

k个总体比率的Marascuilo成对比较方法的临界值

$$CV_{ij} = \sqrt{\chi^2_\alpha}\sqrt{\frac{\bar{p}_i(1-\bar{p}_i)}{n_i}+\frac{\bar{p}_j(1-\bar{p}_j)}{n_j}} \tag{12-3}$$

式中，χ^2_α为显著性水平为α且自由度为$k-1$的χ^2值；$\bar{p}_i$和$\bar{p}_j$为总体i和j的样本比率；n_i和n_j为总体i和j的样本容量。

利用表12-4的χ^2分布表，$k-1=3-1=2$的自由度和显著性水平0.05，我们有$\chi^2_{0.05}=5.991$。现在利用样本比率$\bar{p}_1=0.5520$，$\bar{p}_2=0.6000$和$\bar{p}_3=0.7029$，三个成对比较检验的临界值如下：

雪佛兰羚羊与福特Fusion

$$CV_{12} = \sqrt{5.991}\sqrt{\frac{0.5520(1-0.5520)}{125}+\frac{0.6000(1-0.6000)}{200}} = 0.1380$$

雪佛兰羚羊与本田雅阁

$$CV_{13} = \sqrt{5.991}\sqrt{\frac{0.5520(1-0.5520)}{125}+\frac{0.7029(1-0.7029)}{175}} = 0.1379$$

福特Fusion与本田雅阁

$$CV_{23} = \sqrt{5.991}\sqrt{\frac{0.6000(1-0.6000)}{200}+\frac{0.7029(1-0.7029)}{175}} = 0.1198$$

如果任何成对样本比率之差的绝对值$|\bar{p}_i-\bar{p}_j|$超过其相对应的临界值CV_{ij}，则在显著性水平0.05下，成对之差是显著的，因此我们得出相应的两个总体比率不同。成对比较方法的最后一步汇总在表12-5中。

表12-5　汽车品牌忠诚度研究的成对比较检验

成对比较	$\lvert\bar{p}_i-\bar{p}_j\rvert$	CV_{ij}	如果$\lvert\bar{p}_i-\bar{p}_j\rvert > CV_{ij}$，则显著
雪佛兰羚羊与福特Fusion	0.0480	0.1380	不显著
雪佛兰羚羊与本田雅阁	0.1509	0.1379	显著
福特Fusion与本田雅阁	0.1029	0.1198	不显著

从成对比较方法得出的结论是只有雪佛兰羚羊与本田雅阁的顾客品牌忠诚度之间呈现显著差异。我们的样本结果显示本田雅阁车主表明他们可能再次购买本田雅阁有较大的总体比率，因此，我们能够得出本田雅阁（$\bar{p}_3=0.7029$）比雪佛兰羚羊（$\bar{p}_1=0.5520$）具有较大的顾客品牌忠诚度。

研究结果对比较福特 Fusion 的忠诚度是不确定的。当与雪佛兰羚羊或本田雅阁相比较时，福特 Fusion 没有显示出显著差异。就顾客品牌忠诚度而言，较大的样本或许显示出福特 Fusion 与其他两种品牌的汽车之间有显著差异。对于多重比较方法，显示研究中某些成对比较的显著性，甚至不显示其他成对比较的显著性是很常见的。

注释和评论

1. 在第 10 章中，我们使用标准正态分布和检验统计量 z 进行两个总体比率的假设检验。然而，本节介绍的 χ^2 检验也可以用于进行两个总体比率相等的假设检验。在这两个检验方法下，检验的结果是相同的，而且检验统计量 χ^2 的数值是检验统计量 z 的数值的平方。第 10 章方法的优点是它既可以用于关于两个总体比率的单侧检验，也可以用于双侧检验，而本节的 χ^2 检验只能用于双侧检验。练习 12.6 对两个总体比率相等的假设，给你一个使用 χ^2 检验的机会。
2. 本节中 k 个总体中的每一个都有两种回答结果："是"或"否"。事实上，每一个总体服从二项分布，其参数为 p 为回答"是"的总体比率。当 k 个总体中的每一个有 3 个或更多种可能的回答时，本节 χ^2 方法有一个扩展应用，这时，每一个总体服从多项分布。χ^2 计算的期望频数 e_{ij} 以及检验统计量 χ^2，与式（12-1）和（12-2）相同。唯一不同的是原假设，原假设为对于所有总体回答变量的多项分布是相同的。k 个总体中的每一个都有 r 种回答，检验统计量 χ^2 的自由度为 $(r-1)(k-1)$。练习 12.8 对比较服从多项分布的三个总体，将给你一个使用 χ^2 检验的机会。

练 习

方法

2. 参见练习 1 的观察频数。
   a. 计算每个总体的样本比率。
   b. 在显著性水平 0.05 下，使用多重比较方法确定哪些总体比率差异显著。

应用

4. Benson 制造厂考虑从三家不同的供应商处订购电子部件。就质量而言，有缺陷部件的比率或百分比可能由于供应商的不同而不同。为了评估供应商有缺陷部件的比率，Benson 从每个供应商中抽取一批次的 500 个部件的组成一个样本，其中有缺陷部件和良好部件数量如下：

部件	供应商		
	A	B	C
有缺陷	15	20	40
良好	485	480	460

   a. 建立一个可以用来检验三家供应商提供的有缺陷部件比率相等的假设。
   b. 在显著性水平 0.05 下，进行假设检验。p-值是多少？你的结论如何？
   c. 进行多重比较检验以确定是否存在最佳供应商或是否有供应商因质量差而可能被淘汰。

6. 一家报税公司想比较该公司两个地区办事处的工作质量。[⊖]显示样本报税单中出错申报单数和正确申报单数的观察频数如下：

报税单	地区办事处	
	办事处 1	办事处 2
出错	35	27
正确	215	273

   a. 两个办事处出错申报单的样本比率是多少？
   b. 利用 χ^2 检验方法确定两个办事处的出错率的总体比率之间是否存在显著差异。在显著性水平 0.10 下，检验原假设 $H_0: p_1 = p_2$。p-值是多少？你的结论如何？（注：当有三个或多个总体时，我们通常使用比率相等的 χ^2 检验，但这个例子表明 χ^2 检验同样可以用于检验两个总体比率的相等性。）
   c. 在 10.2 节，使用 z 检验进行了上述检验。检验统计量 χ^2，或检验统计量 z 能都可以用于检验假设。

⊖ 练习 6 说明当假设是关于两个总体比率相等时，可以使用 χ^2 检验。

然而，当我们想要对两个总体比率进行推断时，我们通常首选z检验方法。对于两个总体比率的推断，参见本节末的注释，评价为什么检验统计量z能给使用者提供更多的选项。

8. 某工厂考虑从三个不同的供货商处购买部件。从供货商处收到的部件被分级为有次要缺陷、有主要缺陷和良好。从每个供货商得到的部件样本的检测结果如下表所示。注意对这些数据进行的检验不再是三个总体比率的检验，因为分类型回答变量有三个结果：次要缺陷、主要缺陷和良好。

被检验的部件	供应商		
	A	B	C
次要缺陷	15	13	21
主要缺陷	5	11	5
良好	130	126	124

利用上述数据，进行一个假设检验，以确定三个供货商的缺陷部件的分布是否相同。使用本节所介绍的χ^2检验计算，不同的是具有r行和c列的表将导致检验统计量χ^2的自由度为$(r-1)(c-1)$。在显著性水平0.05下，p-值是多少？你的结论如何？

12.2　独立性检验

χ^2检验的一个重要应用是利用样本数据检验两个分类变量的独立性，为了这个检验，我们从一个总体中抽取样本，并记录两个分类变量的观测值。我们通过对分类变量1和分类变量2的每一对组合统计回答的个数来汇总数据。检验的原假设是两个分类变量独立。因此，这种检验被称为**独立性检验**（test of independence）。我们通过下面的例子来说明这种检验。

啤酒行业协会进行一次调查以确定饮酒者对淡啤酒、普通啤酒和黑啤酒的啤酒偏好。抽取了200名饮酒者组成一个样本，询问样本中每一个人以得到对三种类型啤酒（淡啤酒、普通啤酒和黑啤酒）的偏好。在调查问卷最后，要求回答者提供包括性别（男性或女性）等大量的人口统计学信息。相关的研究问题是三种类型的啤酒偏好是否与饮酒者的性别独立。如果两个分类变量——啤酒偏好和饮酒者性别独立，则啤酒偏好将不依赖于饮酒者性别，且可以期待男性和女性饮酒者对淡啤酒、普通啤酒和黑啤酒的偏好相同。然而，如果检验的结论是两个分类变量不独立，我们将有啤酒偏好与饮酒者的性别有关，或啤酒偏好与饮酒者的性别不独立的证据。因此，我们可以期待男性和女性饮酒者的偏好不同。这样，啤酒厂可以利用这个信息，对男性和女性不同的目标市场采用不同的推销策略。

这个独立性检验的假设如下：

H_0：啤酒偏好与饮酒者性别独立

H_a：啤酒偏好与饮酒者性别不独立

样本数据将汇总在一张二维表中，啤酒偏好（淡啤酒、普通啤酒和黑啤酒）作为一个变量，饮酒者性别（男性和女性）为另一个变量。因为研究的目的是确定男性和女性饮酒者的啤酒偏好之间是否存在差异，因此我们考虑将饮酒者性别作为解释变量。遵循实践中的惯例，将解释变量作为表格中数据的行变量；啤酒偏好是分类回答变量，显示为列变量。研究中200名饮酒者的样本结果汇总在表12-6中。

表12-6　男性与女性饮酒者啤酒偏好的样本资料（观察频数）

		饮酒者性别		
		男性	女性	合计
啤酒偏好	淡啤酒	51	39	90
	普通啤酒	56	21	77
	黑啤酒	25	8	33
	合计	132	68	200

汇总的样本数据基于每个回答者的啤酒偏好和性别的组合。例如，在研究中喜欢淡啤酒的男性有51人，喜欢普通啤酒的男性有56人，等等。现在我们分析表中的数据，并检验啤酒偏好与饮酒者性别独立。

首先，既然我们选择了一个饮酒者的样本，分别对每一个变量进行数据汇总，将得到饮酒者总体特征的一些内在性质。对饮酒者性别这个分类变量，我们看到样本的200人中有132人是男性，这给出饮酒者总体中男性比率的估计值为132/200 = 0.66或66%。同样，饮酒者总体中女性比率的估计值为68/200 = 0.34或34%。因此，男性饮酒者与女性饮酒者人数之比近似为2比1。三种类型啤酒的样本比率或百分比如下：

偏好淡啤酒 $90/200=0.450$ 或 45%

偏好普通啤酒 $77/200=0.385$ 或 38.5%

偏好黑啤酒 $33/200=0.165$ 或 16.5%

综合样本中的所有饮酒者，最常见的啤酒偏好是淡啤酒，最不常见的啤酒偏好是黑啤酒。

现在我们进行χ^2 检验以确定啤酒偏好与饮酒者性别独立。这里的计算结果和所用的公式与12.1 节中所用的χ^2 检验相同。利用表12-6 中第 i 行和第 j 列的观察频数 f_{ij}，在假定啤酒偏好与饮酒者性别独立的条件下，我们计算期望频数 e_{ij}。期望频数遵循 12.1 节同样的逻辑和所用的公式，因此第 i 行和第 j 列的期望频数 e_{ij} 为

$$e_{ij}=\frac{\text{第 } i \text{ 行合计数} \times \text{第 } j \text{ 列合计数}}{\text{样本容量}} \tag{12-4}$$

例如，如果啤酒偏好与饮酒者性别独立，则 $e_{11}=(90\times132)/200=59.40$ 是男性中偏好淡啤酒的期望频数。因此，可以利用式（12-4）确定表12-7 中的其他期望频数。

按照第12.1节χ^2 检验方法，我们用下列表达式计算检验统计量χ^2 的数值：

$$\chi^2=\sum_i\sum_j\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \tag{12-5}$$

对于 r 行和 c 列的表，χ^2 分布的自由度为 $(r-1)\times(c-1)$，且每个单元格的期望频数都大于或等于5。因此，在这个应用中，我们将使用自由度为 $(3-1)\times(2-1)=2$ 的χ^2 分布。计算检验统计量χ^2 的完整步骤汇总在表12-8 中。

表 12-7 在啤酒偏好与饮酒者性别独立时的期望频数

		饮酒者性别		
		男性	女性	合计
啤酒偏好	淡啤酒	59.40	30.60	90
	普通啤酒	50.82	26.18	77
	黑啤酒	21.78	11.22	33
	合计	132	68	200

表 12-8 啤酒偏好与饮酒者性别独立的检验统计量χ^2 的计算

啤酒偏好	性别	观察频数 (f_{ij})	期望频数 (e_{ij})	差 ($f_{ij}-e_{ij}$)	差的平方 $(f_{ij}-e_{ij})^2$	差的平方除以期望频数 $(f_{ij}-e_{ij})^2/e_{ij}$
淡啤酒	男性	51	59.40	-8.40	70.56	1.19
淡啤酒	女性	39	30.60	8.40	70.56	2.31
普通啤酒	男性	56	50.82	5.18	26.83	0.53
普通啤酒	女性	21	26.18	-5.18	26.83	1.02
黑啤酒	男性	25	21.78	3.22	10.37	0.48
黑啤酒	女性	8	11.22	-3.22	10.37	0.92
	合计	200	200			$\chi^2=6.45$

我们可以使用自由度为2 的χ^2 分布的上侧面积和 p-值法，来确定啤酒偏好与饮酒者性别独立的原假设是否被拒绝。利用表12-4 的χ^2 分布表的第2 行，我们有如下信息：

上侧面积	0.10	0.05	0.025	0.01	0.005
χ^2 值 ($df=2$)	4.605	5.991	7.378	9.210	10.597

↑ $\chi^2=6.45$

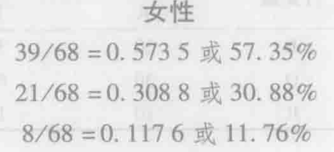
因此，我们看到$\chi^2=6.45$ 的上侧面积介于 0.025 ~ 0.05，所以相对应的上侧面积或 p-值必介于 0.025 ~ 0.05。由于 p-值$\leqslant\alpha=0.05$，我们拒绝 H_0，并得出啤酒偏好与饮酒者性别不独立的结论。另一种方法陈述，研究显示可以期待男性和女性饮酒者的啤酒偏好不同。可以利用附录 F 中提供的 Minitab 或 Excel 程序，对于自由度为2，$\chi^2=6.45$ 对应的 p-值为 0.039 8。

不使用p-值法，我们可以使用临界值法得到相同的结论。对于$\alpha=0.05$和自由度为2，检验统计量χ^2的临界值为$\chi^2=5.991$。上侧检验的拒绝法则变为：

$$\text{如果}\chi^2\geqslant 5.991\text{，则拒绝 }H_0$$

由于$6.45\geqslant5.991$，我们拒绝H_0。因此，我们再次看到p-值法和临界值法得到相同的假设检验结论。

现在我们有啤酒偏好与饮酒者性别不独立的证据，我们需要从数据中得到另外的内在特征，来评估这两个变量之间的关系。一种方法是分别计算男性和女性饮酒者啤酒偏好的概率。这些计算结果如下：



啤酒偏好	男性	女性
淡啤酒	51/132 = 0.386 4 或 38.64%	39/68 = 0.573 5 或 57.35%
普通啤酒	56/132 = 0.424 2 或 42.42%	21/68 = 0.308 8 或 30.88%
黑啤酒	25/132 = 0.189 4 或 18.94%	8/68 = 0.117 6 或 11.76%

男性和女性饮酒者三种类型啤酒的条形图如图12-1所示。

对于啤酒偏好和饮酒者性别之间的关系，你能得到什么样的观察结果？对于样本中的女性饮酒者，最受欢迎的是淡啤酒，占57.35%。对于样本中的男性饮酒者，最受欢迎的是普通啤酒，占42.42%。同时女性饮酒者比男性饮酒者对于淡啤酒有较高偏好，而男性饮酒者中对于普通啤酒和黑啤酒有较高偏好。通过图12-1这样的条形图，数据可视化有助于得到两个分类变量内在关系。

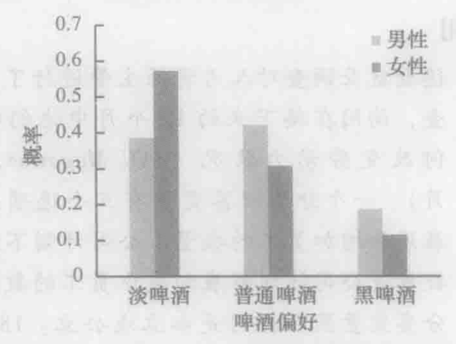


图12-1　啤酒偏好与饮酒者性别的条形图

在我们结束讨论之前，我们汇总独立性检验的步骤如下。

两个分类变量独立性的χ^2检验

1. 建立原假设与备选假设。

$$H_0\text{：两个分类变量独立}$$

$$H_a\text{：两个分类变量不独立}$$

2. 从总体中选择一个随机样本，收集样本中每个个体对两个变量的数据。记录r行和c列的表中的观察频数f_{ij}。

3. 假定原假设为真，计算期望频数e_{ij}。

4. 如果期望频数e_{ij}都大于或等于5[㊀]，计算检验统计量：

$$\chi^2=\sum_i\sum_j\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$$

5. 拒绝法则[㊁]：

$$p\text{-值法：如果 }p\text{-值}\leqslant\alpha\text{，则拒绝 }H_0$$

$$\text{临界值法：如果}\chi^2\geqslant\chi^2_\alpha\text{，则拒绝 }H_0$$

式中，χ^2分布的自由度为$(r-1)(c-1)$，α为检验的显著性水平。

最后，如果独立的原假设被拒绝，像上述例子那样汇总概率有助于分析者确定两个分类变量存在的关联或相关性。

㊀ 期望频数都大于或等于5，对于χ^2检验是有效的。

㊁ χ^2检验也是一个单侧检验，H_0的拒绝域在自由度为$(r-1)(c-1)$的χ^2分布的上侧。

练 习

方法

10. 下表给出了一个容量为240的样本的观察频数。在$\alpha=0.05$下，检验行变量与列变量的独立性。

行变量	列变量		
	A	B	C
P	20	30	20
Q	30	60	25
R	10	15	30

应用

12. 德勤就业调查对人力资源主管进行了一次抽样调查，询问在接下来的12个月中他们公司计划如何改变劳动力状况（*INC. Magazine*，2012年2月）。一个分类回答变量有三个选项：公司计划雇用和增加员工的数量、公司计划不改变员工的数量或公司计划解雇和减少员工的数量。另一个分类变量显示公司是私立或公立。180家公司的样本数据汇总如下：

雇用计划	公司类型	
	私立	公立
增加员工	37	32
不改变	19	34
减少员工	16	42

a. 在显著性水平0.05下，进行独立性检验以接下来的12个月中雇用计划是否与公司类型独立。你的结论如何？

b. 讨论接下来12个月的私立公司和公立公司雇用计划的不同。

14. 汽车质量调查对新车主询问了大量关于他们新近购买汽车的问题（J. D. Power and Associates，2012年3月）。对车主询问的一个问题是对车辆的评价，评价按一般、杰出和不寻常分类。另一个问题询问车主的教育水平，教育水平按近似高中、高中毕业、近似大学和大学毕业分类。假设最近购买汽车的500名车主的样本数据如下：

质量评级	教育水平			
	近似高中	高中毕业	近似大学	大学毕业
一般	35	30	20	60
杰出	45	45	50	90
不寻常	20	25	30	50

a. 在显著性水平0.05下，进行独立性检验以确定新车主的质量评价是否与其教育水平独立。p-值是多少？你的结论如何？

b. 利用一般、杰出和不寻常的总的百分比，评论新车主对他们新近购买汽车的质量的评价。

16. 由于石油价格的上涨，在全世界范围内增加了对替代能源的兴趣。《金融时报》在6个国家所做的一次哈里斯民意调查，评估了民众对各种能源替代形式的态度（Harris Interactive website，2008年2月27日）。下表的数据是人们是否支持和反对建设新核电站的部分调查结果。

民众态度	国家					
	英国	法国	意大利	西班牙	德国	美国
坚决支持	141	161	298	133	128	204
支持大于反对	348	366	309	222	272	326
反对大于支持	381	334	219	311	322	316
坚决反对	217	215	219	443	389	174

a. 这一调查的样本容量是多大？

b. 进行检验假设以确定人们对建设新核电站的态度与国家是否独立。你的结论如何？

c. 根据“坚决支持”和“支持大于反对”回答者的百分比，哪个国家最赞成建设新核电站？哪个国家最不赞成？

18. 在辛迪加的电视节目中关于哪部影片最好，两位主持人经常给人造成强烈争执的印象。每部电影观感按“赞成”“贬斥”或“褒贬皆有”来分类。两位主持人对160部电影的评价结果如下。

主持人A	主持人B		
	贬斥	褒贬皆有	赞成
贬斥	24	8	13
褒贬皆有	8	13	11
称赞	10	9	64

在显著性水平0.01下，用独立性检验分析数据。你的结论如何？

12.3 拟合优度检验

在本节，我们使用χ^2检验来确定一个被抽样的总体是否服从某个特殊的概率分布。首先我们考虑总体服从一个历史的多项概率分布情形，并使用拟合优度检验来确定新的样本数据的总体分布在与历史的分布相比较中，是否显示存在改变。然后我们考虑假设总体服从正态概率分布的情形。在这种情况下，我们利用分布拟合检验来确定样本数据是否显示正态概率分布的假设是适当还是不适当。这两个检验都被称为**拟合优度检验**（goodness of fit test）。

12.3.1 多项概率分布

对于**多项概率分布**（multinomial probability distribution）[㊀]，总体中的每一个个体被分配到三个或多个类中的一个且仅一个。作为一个例子，考虑Scott市场调查公司进行的市场份额研究。在过去的一年中，公司A的市场份额稳定在30%，公司B稳定在50%，公司C稳定在20%。因为每一名顾客可以按照购买这些公司的产品来进行分类，因此我们得到一个三种可能结果的多项概率分布。三个结果中每一种的概率如下：

p_A——顾客购买公司A产品的概率

p_B——顾客购买公司B产品的概率

p_C——顾客购买公司C产品的概率

利用历史的市场份额，我们有多项概率分布[㊁]：$p_A=0.30$，$p_B=0.50$好$p_C=0.20$。

公司C开发了一种“新型改进”产品，以取代当前市场上该公司所售产品。Scott市场调查公司受雇于公司C，目的是判断新产品是否使三家公司的市场份额发生了改变。特别地，Scott市场调查公司将向一个顾客样本介绍公司C的新产品，然后询问顾客对于公司A、公司B及公司C新产品的偏好。基于样本数据，可以使用下列假设检验来确定公司C的新产品是否改变了三家公司的历史市场份额。

$$H_0: p_A = 0.30,\quad p_B = 0.50,\quad p_C = 0.20$$

$$H_a: \text{总体比率不是}\, p_A = 0.30,\quad p_B = 0.50,\quad p_C = 0.20$$

原假设基于市场份额的历史的多项概率分布。如果样本结果导致H_0被拒绝，则Scott市场调查公司有证据表明新产品的引进影响了市场份额。

让我们假定市场调查公司对200名顾客进行了调查，询问每一名顾客对于公司A、公司B及公司C新产品的购买偏好。200份答卷的汇总结果如下。

观察频数		
公司A的产品	公司B的产品	公司C的新产品
48	98	54

现在我们进行拟合优度检验，来确定200名顾客购买偏好与原假设是否相符。与其他χ^2检验一样，拟合优度检验基于样本的观察频数与原假设为真时的期望频数的相比较。因此，下一步就要在假设$H_0: p_A=0.30$，$p_B=0.50$，$p_C=0.20$为真时，计算200名顾客的期望购买偏好，由此得到期望频数如下。

期望频数		
公司A的产品	公司B的产品	公司C的新产品
$200\times0.30=60$	$200\times0.50=100$	$200\times0.20=40$

注意每一类的期望频数是样本容量200与每一类的假设比率的乘积。

现在，拟合优度检验重点关注观察频数与期望频数之间的差异。观察频数与期望频数之间差异的大小，可以借助以下检验统计量χ^2来衡量。

㊀ 多项概率分布是二项概率分布每次试验有三个或多个结果情形的推广。

㊁ 多项概率分布的概率之和等于1。

拟合优度的检验统计量

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{12-6}$$

式中，f_i为第 i 类的观察频数；e_i为第 i 类的期望频数；k 为类别个数。

注意：当所有类别的期望频数都大于或等于5时，检验统计量服从自由度为 $k-1$ 的 χ^2 分布。

让我们继续 Scott 市场调查例子，并利用样本数据来检验假设：多项总体有市场份额的比例 $p_A=0.30$，$p_B=0.50$ 及 $p_C=0.20$ 保持不变。我们在显著性水平 $\alpha=0.05$ 下，用观察频数和期望频数来计算检验统计量的数值。由于期望频数都大于或等于5，在表 12-9 中给出了检验统计量χ^2 计算结果。我们得到$\chi^2=7.34$。

表 12-9 Scott 市场调查公司的市场份额研究中检验统计量χ^2 的计算

类别	假设比率	观察频数 (f_i)	期望频数 (e_i)	差 (f_i-e_i)	差的平方 $(f_i-e_i)^2$	差的平方除以期望频数 $(f_i-e_i)^2/e_i$
公司 A	0.30	48	60	-12	144	2.40
公司 B	0.50	98	100	-2	4	0.04
公司 C	0.20	54	40	14	196	4.90
合计		200				$\chi^2=7.34$

如果观察频数和期望频数之间差异很大，我们将拒绝原假设。因此，拟合优度检验总是上侧检验[⊖]，我们可以用检验统计量的上侧面积和 p-值法来确定原假设是否被拒绝。由于自由度为 $k-1=3-1=2$，利用表 12-4 的χ^2分布表的第 2 行，我们有如下信息：

上侧面积	0.10	0.05	0.025	0.01	0.005
χ^2 数值（$df=2$）	4.605	5.991	7.378	9.210	10.597

↑

$\chi^2=7.34$

检验统计量$\chi^2=7.34$ 介于 5.991 ~ 7.378，因此相应的上侧面积或 p-值必介于 0.025 ~ 0.05。由于 p-值$\leqslant\alpha=0.05$，我们拒绝 H_0，并得出公司 C 引进新产品将改变历史的市场份额的结论。可以利用附录 F 中的 Minitab 或 Excel程序，得到$\chi^2=7.34$ 对应的 p-值为 0.025 5。

不用 p-值法，我们还可以使用临界值法得到相同的结论。由于 $\alpha=0.05$ 和自由度为 2，则检验统计量的临界值为$\chi^2_{0.05}=5.991$，上侧检验的拒绝法则变为

如果$\chi^2\geqslant 5.991$，则拒绝 H_0

由于 7.34 > 5.99，所以我们拒绝 H_0。p-值法和临界值方法，能得到假设检验相同的结论。

现在我们有 C 公司引进新产品将改变三个公司的市场份额的结论，我们有兴趣知道关于市场份额如何变动的更多信息。利用历史的市场份额和样本数据，我们汇总数据如下：

公司	历史市场份额（%）	市场份额的样本数据（%）
A	30	48/200 = 0.24 或 24%
B	50	98/200 = 0.49 或 49%
C	20	54/200 = 0.27 或 27%

⊖ 拟合优度检验总是单侧检验，检验的拒绝域位于χ^2 分布的上侧。

历史市场份额和样本市场份额相比较的条形图如图 12-2 所示。数据可视化方法显示新产品将增加公司 C 的市场份额，与其他两家公司相比，公司 C 市场份额的增加对公司 A 的影响比对公司 B 的影响要大。

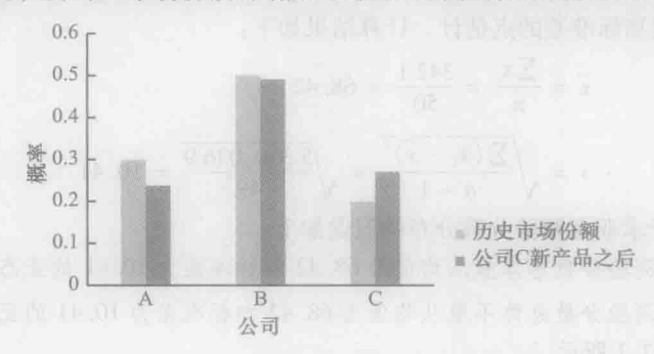


图 12-2　公司 C 引进新产品前后的市场份额的条形图

对假设的多项总体分布进行拟合优度检验的可使用的一般步骤汇总如下。

多项概率分布的拟合优度检验

1. 建立原假设和备选假设。

H_0：总体服从 k 类中每类都有指定的概率的多项概率分布

H_a：总体不服从 k 类中每类都有指定的概率的多项概率分布

2. 选择一个随机样本，并记录每一类别的观察频数 f_i。
3. 假定原假设为真，由此确定每一类别的期望频数 e_i，e_i 等于样本容量与各类概率的乘积。
4. 如果每一类的期望频数 e_i 都大于或等于 5，计算检验统计量的数值。

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

5. 拒绝法则：

p-值法：如果 p-值 $\leqslant \alpha$，则拒绝 H_0

临界值法：如果 $\chi^2 \geqslant \chi^2_\alpha$，则拒绝 H_0

式中，α 为检验的显著性水平；$k-1$ 为自由度。

12.3.2　正态分布

正态分布拟合优度检验也是基于 χ^2 分布的应用。具体地，在总体服从正态分布的假设下，将样本数据的若干类别的观察频数同期望频数进行比较。因为正态分布是连续型的，我们必须修正定义类别的方式以及计算期望频数的方法。让我们通过考察表 12-10 中 Chemline 公司求职者的测验数据，来阐述正态分布的拟合优度检验。

表 12-10　Chemline 公司随机选取 50 名求职者的能力测验分数

71	66	61	65	54	93	60	86	70	70	73	73
55	63	56	62	76	54	82	79	76	68	53	58
85	80	56	61	61	64	65	62	90	69	76	79
77	54	64	74	65	65	61	56	63	80	56	71
79	84										

Chemline 公司每年为其遍布美国的 4 家工厂招聘大约 400 名新雇员。人事部主管想知道测验分数总体是否呈现正态分布。如果呈现正态分布，则可以根据这个分布评估具体分数，即可以迅速判断出前 20% 人的分数、后

40%人的分数，等等。因此，我们试图检验测验分数总体服从正态分布的原假设。

首先，我们利用表 12-10 中的数据估计原假设中要考虑的正态分布的均值和标准差。我们用样本均值 $\bar{x}$ 和样本标准差 s 作为正态分布均值和标准差的点估计。计算结果如下：

$$\bar{x} = \frac{\sum x_i}{n} = \frac{342\,1}{50} = 68.42$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{5\,310.036\,9}{49}} = 10.41$$

利用这些数值，我们提出关于求职者测验分数分布的假设如下：

H_0：测验分数总体服从均值为 68.42 和标准差为 10.41 的正态分布

H_a：测验分数总体不服从均值为 68.42 和标准差为 10.41 的正态分布

所假设的正态分布如图 12-3 所示。

由于正态分布是连续型的，我们必须采取不同的方法来定义类别。我们需要利用测验分数的区间来定义类别。

回顾以上关于每个区间或类别中期望频数至少为 5 的法则。我们定义测验分数的类别时，也必须使每一类的期望频数至少为 5。由于样本容量为 50，一种方法是将正态分布划分为 10 个等概率区间（见图 12-4）。由于样本容量为 50，我们可以期望每个区间或类别中有 5 个结果，关于期望频数至少为 5 的法则也满足了。[⊖]

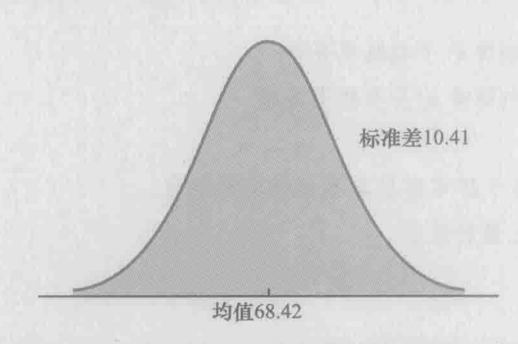


图 12-3 Chemline 公司求职者测验分数假设的正态分布

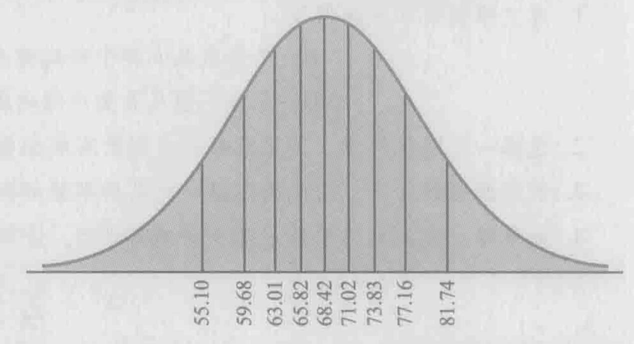


图 12-4 Chemline 公司例子的 10 个等概率区间的正态分布

注：每个区间的概率为 0.1。

我们进一步考察计算类别边界的方法。当假定为正态概率分布时，标准正态概率表可以用于确定这些边界。首先考虑最低 10% 的测验分数的分界值。根据标准正态分布表，我们得到该测验分数对应的 z 值为 -1.28。因此，分数 $x = 68.42 - 1.28 \times 10.41 = 55.10$ 是最低 10% 的分界值。对于最低 20% 的情形，我们得到 $z = -0.84$，于是 $x = 68.42 - 0.84 \times 10.41 = 59.68$。用这种方法处理整个正态分布，得到下列测验分数值。

百分比（%）	z 值	测验分数
10	−1.28	68.42 − 1.28 × 10.41 = 55.10
20	−0.84	68.42 − 0.84 × 10.41 = 59.68
30	−0.52	68.42 − 0.52 × 10.41 = 63.01
40	−0.25	68.42 − 0.25 × 10.41 = 65.82
50	0.00	68.42 + 0.00 × 10.41 = 68.42
60	+0.25	68.42 + 0.25 × 10.41 = 71.02
70	+0.52	68.42 + 0.52 × 10.41 = 73.83
80	+0.84	68.42 + 0.84 × 10.41 = 77.16
90	+1.28	68.42 + 1.28 × 10.41 = 81.74

⊖ 对于连续型概率分布，依据使每个区间中的期望频数至少为 5 的法则来建立区间。

这些区间的分界值或边界点已标注在图 12-4 中。

现在测验分数的类别或区间已定义好了，且已知每一类的期望频数为 5，我们可以回到表 12-10 的样本数据中来确定这些类别中的观察频数。由此得出表 12-11 的结果。

表 12-11 Chemline 公司求职者测验分数的观察频数和期望频数

测验分数区间	观察频数（f_i）	期望频数（e_i）
55.10 以下	5	5
55.10 ~ 59.68	5	5
59.68 ~ 63.01	9	5
63.01 ~ 65.82	6	5
65.82 ~ 68.42	2	5
68.42 ~ 71.02	5	5
71.02 ~ 73.83	2	5
73.83 ~ 77.16	5	5
77.16 ~ 81.74	5	5
81.74 以上	6	5
合计	50	50

利用表 12-11 中的计算结果，拟合优度检验的计算过程与前面完全一样。即我们通过计算 χ^2 的数值来比较观察频数与期望频数。检验统计量 χ^2 的必要计算结果列在表 12-12 中，我们得到检验统计量的数值为 $\chi^2=7.2$。

为了确定 χ^2 的数值 7.2 是否大到足以拒绝 H_0，我们需要参考适当的 χ^2 分布表。利用拟合优度检验中计算自由度的法则，我们得到自由度为 $k-p-1=10-2-1=7$，这里有 $k=10$ 个类别以及由样本估计的 $p=2$ 个参数（均值与标准差）。[⊖]

表 12-12 Chemline 公司求职者例子的统计量 χ^2 的计算

测验分数区间	观察频数（f_i）	期望频数（e_i）	差（f_i-e_i）	差的平方（f_i-e_i）^{2}	差的平方除以期望频数（f_i-e_i）$^2/e_i$
55.10 以下	5	5	0	0	0.0
55.10 ~ 59.68	5	5	0	0	0.0
59.68 ~ 63.01	9	5	4	16	3.2
63.01 ~ 65.82	6	5	1	1	0.2
65.82 ~ 68.42	2	5	−3	9	1.8
68.42 ~ 71.02	5	5	0	0	0.0
71.02 ~ 73.83	2	5	−3	9	1.8
73.83 ~ 77.16	5	5	0	0	0.0
77.16 ~ 81.74	5	5	0	0	0.0
81.74 以上	6	5	1	1	0.2
合计	50	50			$\chi^2=7.2$

假设我们在显著性水平 0.10 下，检验测验分数服从正态分布的原假设。为了检验这个假设，我们需要根据自由度为 7 的 χ^2 分布确定 $\chi^2=7.2$ 的上侧面积，从而确定 p-值。利用表 12-4 的第 7 行，我们看到 $\chi^2=7.2$ 的上侧面积超过 0.10。因此，我们知道 p-值大于 0.10。利用书后附录 F 中的 Minitab 或 Excel 程序得到 $\chi^2=7.2$ 对应的 p-值为 0.408 4。由于 p-值 >0.10，因此，Chemline 公司求职者测验分数服从正态分布的假设不能被拒绝。应用正态分布可以有助于解释测验分数。正态分布的拟合优度检验的步骤总结如下。

正态分布的拟合优度检验

1. 建立原假设和备选假设

H_0：总体服从正态概率分布

H_a：总体不服从正态概率分布

2. 抽取一个随机样本，并且

a. 计算样本均值和样本标准差。

⊖ 估计正态分布中的两个参数将导致 χ^2 检验的自由度减少 2 个。

b. 定义k个取值区间使得每个区间中的期望频数至少为5。使用等概率区间是很好的方法。

c. 记录每个确定好的区间的数据值的观察频数f_i。

3. 对于步骤2（b）中确定的每个区间，计算发生次数的期望频数e_i，即样本容量与正态随机变量落入每个区间的概率的乘积。

4. 计算检验统计量

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

5. 拒绝法则：

p-值法：如果p-值$\leqslant \alpha$，则拒绝H_0

临界值法：如果$\chi^2 \geqslant \chi^2_\alpha$，则拒绝$H_0$

式中，α为显著性水平；自由度为$k-p-1$，其中p是有样本估计的分布参数的个数。在步骤2（a）中，利用样本估计了均值和标准差。因此，$p=2$，自由度为$k-2-1=k-3$。

练习

方法

☆20. 以下数据被认为是来自一个正态分布。在$\alpha=0.05$下，进行拟合优度检验，以检验这个假设。

17 23 22 24 19 23 18 22
20 13 11 21 18 20 21
21 18 15 24 23 23 43 29
27 26 30 28 33 23 29

应用

22. 玛氏公司生产一种世界上最受欢迎的糖果品牌——M&M。牛奶巧克力糖果有各种颜色，其中包括蓝色、褐色、绿色、橙色、红色和黄色（M&M website，2012年3月）。颜色的总比率是：蓝色0.24，褐色0.13，绿色0.20，橙色0.16，红色0.13和黄色0.14。在一次抽样研究中，打开几袋M&M牛奶巧克力，得到颜色数量结果如下。

蓝色	褐色	绿色	橙色	红色	黄色
105	72	89	84	70	80

在显著性水平0.05下，用样本数据检验假设：颜色的总比率如前所述。你的结论如何？

24. 美国公路交通安全局报道了一周中每天发生交通事故的百分比（《时代周刊》，2012年3月12日）。假设420次交通事故的样本中提供了如下数据：

星期日	星期一	星期二	星期三	星期四	星期五	星期六
66	50	53	47	55	69	80

a. 在显著性水平0.05下，进行假设检验以确定一周中每天发生交通事故的比率是否相同。你的结论如何？

b. 计算一周中每天发生交通事故的比率。哪天发生交通事故的比率最高？这合理吗？请讨论。

26. 某种产品的每周需求量被认为服从正态分布。在$\alpha=0.10$下，利用拟合优度检验以及下列数据检验这一假设。样本均值为24.5，样本标准差为3。

18 20 22 27 22 25 22 27 25
24 26 23 20 24 26
27 25 19 21 25 26 25 31 29
25 25 28 26 28 24

小结

本章我们介绍了对下列应用的假设检验：

1. 检验三个或多个总体比率的相等性。
2. 检验两个分类变量的独立性。
3. 检验一个总体的概率分布是否服从一个特殊的历史的或理论的概率分布。

所有这些检验都适用于分类变量，且所有的检验都利用检验统计量χ^2，该统计量基于观察频数与期望频数之间的差异。在每一种情形中，在原假设为真时，

计算期望频数，这些χ^2检验都是上侧检验。观察频数与期望频数之间差异较大，将导致检验统计量χ^2的数值较大，则表明原假设应该被拒绝。

三个或多个总体比率相等性的检验基于从每一个总体独立抽取的随机样本。对每一个总体的两类回答，样本数据提供了每一类的个数。原假设是所有总体比率相等，原假设的拒绝域支持所有总体比率不全相等的结论。

两个分类变量的独立性检验利用来自总体的一个样本，数据显示两个分类变量的每一个组合的个数。原假设是两个变量独立，这个检验被称为独立性检验。如果原假设被拒绝，则有两个变量之间相关或依赖的统计证据。

利用拟合优度检验，来检验一个总体服从某个特殊的历史或理论概论分布的假设。我们对服从多项概率分布和正态概率分布阐述了应用。由于正态概率分布适用于连续型数据，因此对拟合优度检验需要的分类变量，设置数据值的区间来定义类别。

关键术语

Marascuilo procedure　Marascuilo 方法　一种用于检验成对总体比率之间差异显著性的多重比较方法。当所有总体比率相等的原假设被拒绝时，这个检验有助于确定成对总体比率之间的差异。

test of independence　独立性检验　一种用于检验两个分类变量独立性的χ^2检验。如果独立性假设被拒绝，则可以得出两个分类变量相关或依赖的结论。

goodness of fit test　拟合优度检验　一种用于检验一个总体概率分布服从一个特殊的历史或理论分布的χ^2检验。多项概率分布和正态概率分布用来说明这种检验。

multinomial probability distribution　多项概率分布　每一个结果属于三个或多个类别中的一个且仅一个的一种概率分布。多项概率分布将二项概率分布由每次试验的两个结果推广到三个或多个结果。

重要公式

在 H_0 为真的条件下的期望频数

$$e_{ij} = \frac{\text{第 } i \text{ 行合计数} \times \text{第 } j \text{ 列合计数}}{\text{样本容量}} \tag{12-1}$$

检验统计量χ^2

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \tag{12-2}$$

Marascuilo 成对比较方法的临界值

$$CV_{ij} = \sqrt{\chi^2_\alpha}\sqrt{\frac{\bar{p}_i(1-\bar{p}_i)}{n_i} + \frac{\bar{p}_j(1-\bar{p}_j)}{n_j}} \tag{12-3}$$

拟合优度的检验统计量χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{12-6}$$

补充练习

28. 菲尼克斯国际营销组织认定康涅狄格州的布里奇波特、新墨西哥州的洛斯阿拉莫斯、佛罗里达州的那不勒斯和华盛顿特区为美国百万富翁比率最高的四个城市（《今日美国》，2011年12月7日）。从四个城市中的每一个分别抽取个人样本，与研究相关的百万富翁数量如下：

是否百万富翁	城市			
	布里奇波特	洛斯阿拉莫斯	那不勒斯	华盛顿特区
是	44	35	36	34
否	456	265	364	366

a. 每个城市的百万富翁比率的估计值分别是多少？

b. 在显著性水平0.05下，对四个城市百万富翁总体比率的相等性进行检验。p-值是多少？你的结论如何？

30. Pew研究中心进行的一项调查，询问人们更适合于在节奏较慢还是较快的地区生活（《今日美国》，2009年2月13日）。调查还询问了回答者的性别。考虑如下样本数据。

喜欢的生活节奏	性别	
	男性	女性
较慢	230	218
无偏好	20	24
较快	90	48

a. 在显著性水平0.05下，喜欢的生活节奏与性别是否独立？p-值是多少？你的结论如何？

b. 讨论男性与女性偏好之间的差异。

32. 弗吉尼亚州两个县紧急救护车呼叫次数的数据如下。

一个县是城市型县，另一个县是乡村型县。在过去两年中，471 次救护车呼叫的一个样本给出了每次紧急呼叫的县的类型和星期几。样本数据如下：

县的类型	星期						
	日	一	二	三	四	五	六
城市型县	61	48	50	55	63	73	43
乡村型县	7	9	16	13	9	14	10

在显著性水平 0.05 下，检验县的类型与星期的独立性。p-值是多少？你的结论如何？

34. 一门大学课程期末考试成绩的随机样本如下。

55	85	72	99	48	71	88	70	59	
98	80	74	93	85	74				
82	90	71	83	60	95	77	84	73	
63	72	95	79	51	85				
76	81	78	65	75	87	86	70	80	64

在显著性水平 0.05 下，检验考试成绩的总体分布为正态分布的假设是否应该被拒绝。

案例　两党议程变更

在 Zogby International 为 *Democrat and Chronicle* 进行的一次研究中，对 700 多位纽约人进行了民意测验，以判断纽约州政府是否正常运作。受访者被问及的问题涉及州议员的减薪、对说客的限制、州议员任期的限制、以及州公民是否能够将事项直接交给州议员进行投票表决。关于一些改革建议的成效，跨越所有人口的和政治的界限，得到了选民的广泛支持。

假定，对居住在纽约州西部地区的 100 名个人进行了一次跟进调查。每位受访者的党派（民主党、独立党、共和党），以及他们对下列三个问题的回答被记录下来。

1. 因为每天的州预算都滞后，州议员是否应该减薪？

   是____　否____

2. 是否应该对说客加以更多的限制？

   是____　否____

3. 是否应该对州议员需要服务的任期有一个固定的年限？

   是____　否____

回答用 1 代表“是”，用 2 代表“否”。完整的数据集可以提供给我们使用，存在光盘中名为 NYReform 的文件中。

管理报告

1. 利用描述统计汇总该研究中的数据。对调查涉及的三个问题中的每一个问题，其回答（是与否）与党派之间的独立性，你有何初步结论？
2. 对于问题 1，在显著性水平 $\alpha=0.05$ 下，检验受访者的回答（是与否）与党派的独立性。
3. 对于问题 2，在显著性水平 $\alpha=0.05$ 下，检验受访者的回答（是与否）与党派的独立性。
4. 对于问题 3，在显著性水平 $\alpha=0.05$ 下，检验受访者的回答（是与否）与党派的独立性。
5. 跨越所有的政治界限，改革建议是否得到了选民的广泛支持？请解释。

第13章 实验设计与方差分析

CHAPTER 13

实践中的统计

Burke 市场营销服务公司[⊖]

俄亥俄州，辛辛那提

Burke 市场营销服务公司是工业界最富有经验的市场研究机构之一。与世界上任何其他市场研究公司相比，Burke 每天都有更多的提议、更多的项目出笼。由于有当前美国最先进的科学技术的支撑，Burke 具备非常广泛的研究能力，可以对几乎所有市场营销问题提供答案。

在一项研究中，Burke 受聘于一家公司来为儿童干谷类食品的潜在新品种作出评价。为了保守商业秘密，我们称这家谷类食品制造商为 Anon 公司。Anon 产品开发者认为可能改善谷类食品味道的四个关键因素为：

1. 谷类食品中小麦与玉米的比例；
2. 甜味剂的类型：食糖、蜂蜜或人工增甜剂；
3. 有无果味香精；
4. 加工时间的长短。

Burke 设计了一个用于确定这四个因素对谷类食品味道将会产生什么影响的实验。例如，一种测试的谷类食品是在某个特定的小麦与玉米的比例、甜味剂为食糖、加入果味香精和短加工时间条件下制成的；另一种测试的谷类食品是在小麦与玉米的不同比例，但其他三个因素相同的条件下制成的，等等。由参加测试的几组儿童品尝这些谷类食品并且对每种食品的味道进行评价。

方差分析是一种统计方法，我们使用这种方法来研究儿童品尝谷类食品的味道得到的数据。下面显示的是分析结果：

1. 谷类食品的成分及甜味剂的类型对味道评价的影响很大。
2. 果味香精事实上破坏了谷类食品的味道。
3. 加工时间对谷类食品的味道没有影响。

这些信息帮助 Anon 公司识别出了可能生产出最佳味道谷类食品的因素。

Burke 进行的实验设计及随后的方差分析对生产谷类食品的设计方案很有裨益。在本章里，我们将看到这些方法是如何实现的。

在第 1 章我们提到，统计研究可以分为实验性研究与观测性研究两类。在实验性统计研究中，数据是通过实验产生的。一项实验首先要从确定一个我们感兴趣的变量开始。然后确定并控制一个或多个其他变量，这些其他变量与我们感兴趣的变量是相关的；与此同时，收集这些变量如何影响我们感兴趣的那一个变量的数据。

在观测性研究中，我们经常是通过抽样调查，而不是通过控制一项实验来获取数据。一些好的设计原则仍然会得到使用，但严格控制一项实验性统计研究往往是不可能的。例如，在一项有关吸烟与肺癌之间关系的研究中，研究人员不可能为实验性研究的对象指定其是否有吸烟嗜好。研究人员仅限于简单地观察吸烟对那些曾经吸烟的人的影响，以及不吸烟对那些已经不吸烟的人的影响。

在本章中，我们介绍三种类型的实验设计：完全随机化设计、随机化区组设计以及析因实验。对于每一种实验设计，我们将要说明，方差分析（ANOVA）的统计方法如何能用于现有数据的分析。我们也可使用 ANOVA 来分析通过观测性研究得到的数据。例如，我们将会看到，当数据是通过一项观测性研究得到时，用于完全随机化的实验设计的 ANOVA 方法也适用检验三个或三个以上总体均值的相等问题。在第 14 章我们将会看到，在分析涉及实验性研究与观测性研究两种类型数据的回归研究结果时，ANOVA 方法起到了关键的作用。[⊜]

⊖ 作者感谢 Burke 市场营销服务公司的罗纳德·塔特姆博士，他为“实践中的统计”提供了本案例。

⊜ 罗纳德 A. 费雪（Ronald Alymer Fisher，1890～1962）创立了被称为实验设计的统计学分支。他不仅在统计学领域卓有成就，而且还是遗传学领域的著名科学家。

在 13.1 节，我们将介绍实验性研究的基本原理，并且将说明，如何将这些基本原理应用到完全随机化的设计中。然后，在 13.2 节我们将要说明，如何应用 ANOVA 来分析完全随机化实验设计的数据。在后面的几节中，我们将讨论多重比较方法和另外两个有广泛应用的实验设计：随机化区组设计和析因实验。

13.1　实验设计和方差分析简介

作为实验性统计研究的例子，我们考虑 Chemitech 公司遇到的问题。Chemitech 公司开发了一种新的城市供水过滤系统。新过滤系统的部件需要从几家供应商处购买，然后由 Chemitech 公司设在南加州哥伦比亚市的工厂装配这些部件。公司的工程部负责确定新过滤系统的最佳装配方法。考虑了各种可能的装配方法后，工程部将范围缩小至三种方法：方法 A、方法 B 及方法 C。这些方法在新过滤系统装配步骤的顺序上有所不同。Chemitech 公司的管理人员希望确定：哪种装配方法能使每周生产的过滤系统的数量最多。[㊀]

在 Chemitech 公司的实验中，装配方法是独立变量或**因子**（factor）。因为对应于这个因子有三种装配方法，所以我们说这一实验有三个处理；每个**处理**（treatment）对应于三种装配方法中的一种。Chemitech 公司的问题是一个**单因子实验**（single-factor experiment）的实例；该问题只涉及一个定性因子（装配方法）。更为复杂的实验可能由多个因子组成；其中有些因子可能是定性的，有些因子可能是定量的。

三种装配方法或处理确定了 Chemitech 公司实验的三个总体。一个总体是使用装配方法 A 的全体工人，第二个总体是使用装配方法 B 的全体工人，第三个总体是使用装配方法 C 的全体工人。注意：对每个总体，因变量或**响应变量**（response variable）是每周装配的过滤系统的数量，并且该实验的主要统计目的是确定：三个总体（三种方法）每周所生产的过滤系统的平均数量是否相同。

假设从 Chemitech 公司的生产车间的全体装配工人中抽取了三名工人组成一个随机样本。用实验设计的术语，三名随机抽取的工人是**实验单元**（experiment units）。我们将在 Chemitech 公司的问题中使用的实验设计称为**完全随机化设计**（completely randomized design）。这种类型的设计要求将每一种装配方法或处理随机地指派给一个实验单元或一名工人。例如，方法 A 可能被随机地指派给第二名工人，方法 B 指派给第一名工人，方法 C 指派给第三名工人。如同本例所解释的那样，随机化的概念是所有实验设计的一个重要原则。[㊁]

注意：这个实验对每个处理只会得到一个装配好的过滤系统的测度或数量。对于每种装配方法，为了得到更多的数据，我们必须重复或复制基本的实验过程。例如，假设我们不是只随机抽取 3 名工人，而是 15 名工人，然后对每一个处理随机地指派 5 名工人。因为每种装配方法都指派给 5 名工人，因此我们说得到了 5 个复制。复制的过程是实验设计的另一个重要原则。图 13-1 显示了 Chemitech 公司实验的完全随机化设计。

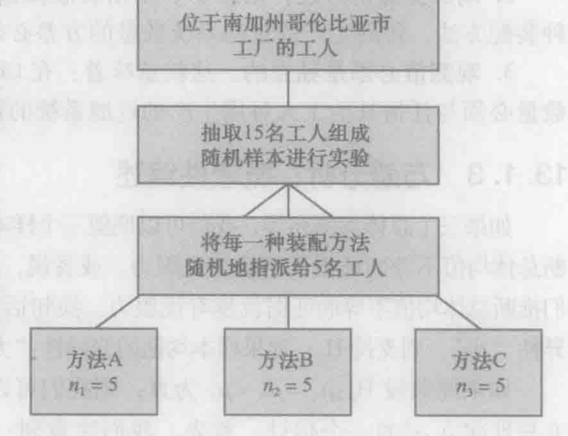


图 13-1　评价 Chemitech 公司装配方法实验的完全随机化设计

13.1.1　数据收集

一旦我们对实验设计感到满意，我们将进行收集和分析数据的工作。在 Chemitech 公司的例子中，工人们将得到如何按照指派给他们的装配方法工作的培训，然后使用这种方法开始装配新的过滤系统。在指派装配方法及培训工作都已经完成后，一周内每名工人装配的过滤系统的数量

㊀ 因果关系在观测性研究中是很难证实的，但是在实验性研究中很容易证实因果关系。

㊁ 随机化是将处理随机地指派给实验单元的过程。在罗纳德 A. 费雪以前，处理是以系统的方式或主观的方式被指派的。

如表13-1所示。每一种装配方法所生产的过滤系统的样本均值、样本方差和样本标准差也在表中给出。于是，使用装配方法A生产的过滤系统的样本均值是62；使用装配方法B生产的过滤系统的样本均值是66；使用装配方法C生产的过滤系统的样本均值是52。从这些数据可以看出，装配方法B似乎比其他装配方法有更高的劳动生产率。

表13-1 15名工人生产的过滤系统的数量

	方法		
	A	B	C
	58	58	48
	64	69	57
	55	71	59
	66	64	47
	67	68	49
样本均值	62	66	52
样本方差	27.5	26.5	31.0
样本标准差	5.244	5.148	5.568

真正的问题是，观察到的三个样本均值之间的差异是否足够大，以致使我们能够得出结论：对应于三种装配方法的总体均值是不同的。为了用统计术语来描述这一问题，我们引入下列记号。

μ_1——使用装配方法A每周生产的过滤系统的数量

μ_2——使用装配方法B每周生产的过滤系统的数量

μ_3——使用装配方法C每周生产的过滤系统的数量

尽管我们根本不可能知道μ_1，μ_2和μ_3的实际数值，但我们还是试图用样本均值来检验下面的假设。

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_a: \text{总体均值不全相等}$$

正如我们很快将要证明的那样，利用方差分析（ANOVA）这一统计方法可以确定，在三个样本均值之间观察到的差异是否足够大到可以拒绝H_0。[⊖]

13.1.2 方差分析的假定

应用方差分析需要三个假定。

1. **对每个总体，响应变量服从正态分布**[⊜]。这就意味着：在Chemitech公司的实验中，对于每一种装配方法，每周生产的过滤系统的数量（响应变量）必须服从正态分布。

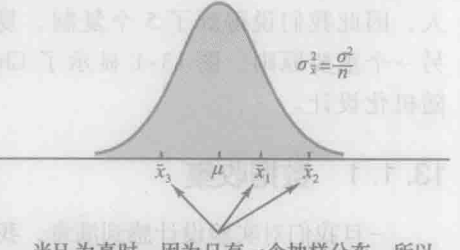
2. **响应变量的方差，记为σ^2，对所有总体都是相同的**。这就意味着：在Chemitech公司的实验中，对于每一种装配方法，每周生产的过滤系统数量的方差必须是相同的。

3. **观测值必须是独立的**。这就意味着：在Chemitech公司的实验中，对于每名工人，每周生产的过滤系统的数量必须与任何其他工人每周生产的过滤系统的数量独立。

13.1.3 方差分析：概念性综述

如果三个总体均值相等，我们可以期望三个样本均值彼此之间很接近。事实上，三个样本均值相互越接近，我们推断总体均值不等的证据就越缺乏说服力。或者说，样本均值差异越大，我们推断总体均值不等的证据就越有说服力。换句话说，如果样本均值的变异性“小”，则支持H_0；如果样本均值的变异性“大”，则支持H_a。

如果原假设$H_0: \mu_1 = \mu_2 = \mu_3$为真，则我们可以利用样本均值之间的变异性建立$\sigma^2$的一个估计。首先，我们注意到：如果方差分析的假设成立并且原假设为真，则每一个样本都是来自均值为μ，方差为σ^2的同一正态分布。在第7章我们曾讲过，来自正态总体的容量为n的一个简单随机样本的样本均值$\bar{x}$的抽样分布仍然服从正态分布，其均值为μ，方差为σ^2/n。图13-2用图示说明了这一抽样分布。



当H_0为真时，因为只有一个抽样分布，所以几个样本均值彼此之间是“相互接近的”

图13-2 H_0为真时$\bar{x}$的抽样分布

⊖ 若H_0被拒绝，我们不能得出总体均值全不相等的结论。拒绝H_0意味着至少两个总体的均值有不同的数值。

⊜ 如果样本容量相等，方差分析对于违背总体服从正态分布的假定不敏感。

于是，如果原假设为真，我们可以把由表 13-1 得到的三个样本均值 $\bar{x}_1=62$，$\bar{x}_2=66$，$\bar{x}_3=52$ 中的每一个，都认为是从图 13-2 所示的抽样分布中随机抽取的数值。在这种情况下，三个样本均值 $\bar{x}_1$、$\bar{x}_2$ 和 $\bar{x}_3$ 的均值与方差可以用来估计该抽样分布的均值与方差。例如，在 Chemitech 公司实验的例子中，当样本容量相等时，$\bar{x}$ 的抽样分布的均值的最佳估计是三个样本均值的均值或算术平均数。在 Chemitech 公司实验的例子中，$\bar{x}$ 抽样分布的均值的一个估计值是 $(62+66+52)/3=60$。我们称该估计值为总样本均值。$\bar{x}$ 抽样分布的方差 $\sigma_{\bar{x}}^2$ 的估计可以由三个样本均值的方差给出。

$$s_{\bar{x}}^2=\frac{(62-60)^2+(66-60)^2+(52-60)^2}{3-1}=\frac{104}{2}=52$$

由 $\sigma_{\bar{x}}^2=\sigma^2/n$，解得

$$\sigma^2=n\sigma_{\bar{x}}^2$$

于是

$$\sigma^2 \text{ 的估计值} = n\times(\sigma_{\bar{x}}^2 \text{ 的估计量}) = ns_{\bar{x}}^2 = 5\times 52 = 260$$

所得结果 $ns_{\bar{x}}^2=260$ 称为 σ^2 的处理间估计。

σ^2 的处理间估计的根据是：原假设为真。在这种情形下，每个样本都来自同一个总体，并且 $\bar{x}$ 只有一个抽样分布。为了说明 H_0 为假时发生了什么情况，假定总体均值全不相同。注意，由于三个样本分别来自均值不同的三个正态分布，因此将导致有三个不同的抽样分布。图 13-3 表明在这种情形下，样本均值彼此之间不再像 H_0 为真时那样接近了。于是，$s_{\bar{x}}^2$ 将会变得比较大，从而使得 σ^2 的处理间估计也变得比较大。一般地，当总体均值不相等时，处理间估计将会高估总体方差 σ^2。

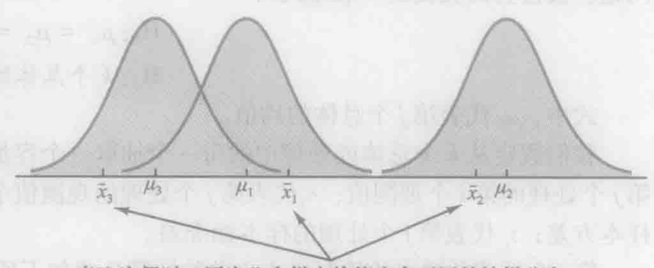


图 13-3　H_0 为假时 $\bar{x}$ 的抽样分布

每个样本内部的变异也会对我们得到的方差分析的结论产生影响。当我们从每一个总体中抽取一个随机样本时，每个样本方差都给出了 σ^2 的一个无偏估计。因此，我们可以将 σ^2 的个别估计组合或合并成一个总的估计。用这种方法得到的 σ^2 的估计称为 σ^2 的合并估计或处理内估计。因为每个样本方差给出的 σ^2 的估计仅以每个样本内部的变异为依据，因此，σ^2 的处理内估计不受总体均值是否相等的影响。当样本容量相等时，σ^2 的处理内估计可以通过计算个别样本方差的算术平均值得到。对于 Chemitech 公司实验的例子，我们有

$$\sigma^2 \text{的处理内估计}=\frac{27.5+26.5+31.0}{3}=\frac{85}{3}=28.33$$

在 Chemitech 公司实验的例子中，σ^2 的处理间估计（260）远大于 σ^2 的处理内估计（28.33）。事实上，这两个估计量的比值为 260/28.33 = 9.18。但是，我们回想起：只有当原假设为真时，处理间估计方法才是总体方差 σ^2 的一个好的估计量；如果原假设为假，处理间估计方法将高估总体方差 σ^2。不过在这两种情形下，处理内估计都是总体方差 σ^2 的一个好的估计量。因此，如果原假设为真，则两个估计量应该很接近，并且它们的比值接近于 1。如果原假设为假，则处理间估计将大于处理内估计，并且它们的比值也将是大的。在下一节我们将说明，为了拒绝 H_0，这个比值必须达到多大。

总的说来，ANOVA 背后的逻辑是以共同总体方差 σ^2 的两个独立的估计量为基础。σ^2 的一个估计量是以样本均值它们自己之间的变异性为依据，σ^2 的另一个估计量是以每个样本内部数据的变异性为依据。通过比较 σ^2 的这两个估计量，我们就能够确定总体均值是否相等。

注释和评论

1. 实验设计中的随机化与观测性研究中的概率抽样相类似。
2. 在许多医学实验中，潜在的偏差通过使用双盲（double-blind）的实验设计被排除。在这样的设计中，无

论是应用处理的医生还是受试对象，都不知道应用的是哪一种处理。许多其他类型的实验也可以从这种类型的设计中受益。

3. 对于一个完全随机化实验设计，我们在本节中给出了如何应用方差分析来检验 k 个总体均值相等的一个概念性的叙述。我们将看到，对于观测性或非实验性研究，也可以用同样的程序来检验 k 个总体均值相等的问题。
4. 在10.1节和10.2节中我们已经介绍了检验两个总体均值相等的假设的统计方法。ANOVA也可以用来检验两个总体均值相等的假设。但在实践中，除非在处理三个或三个以上总体的均值问题时，通常不使用方差分析方法。

13.2 方差分析和完全随机化实验设计

在本节中，我们将说明，对于一个完全随机化实验设计，如何应用方差分析来检验 k 个总体均值是否相等的问题。被检验的假设的一般形式为

$$\mathrm{H}_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$\mathrm{H}_a: k \text{ 个总体的均值不全相等}$$

式中，μ_j 代表第 j 个总体的均值。

我们假定从 k 个总体或处理中的每一个抽取一个容量为 n_j 的简单随机样本。对于得到的样本数据，令 x_{ij} 代表第 j 个处理的第 i 个观测值；n_j 代表第 j 个处理的观测值个数；$\bar{x}_j$ 代表第 j 个处理的样本均值；s_j^2 代表第 j 个处理的样本方差；s_j 代表第 j 个处理的样本标准差。

第 j 个处理的样本均值与样本方差的计算公式如下所示。

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \tag{13-1}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1} \tag{13-2}$$

总样本均值，记为 $\bar{\bar{x}}$，等于所有观测值之和除以观测值的总个数。即

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T} \tag{13-3}$$

式中

$$n_T = n_1 + n_2 + \cdots + n_k \tag{13-4}$$

若每个样本的容量是相等的，都为 n，则 $n_T = kn$；在这种情形下，式（13-3）简化为

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{kn} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}/n}{k} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k} \tag{13-5}$$

换句话说，只要样本容量全相等，总样本均值恰好是 k 个样本均值的算术平均数。

在Chemitech公司实验的例子中，因为每个样本都是由 $n=5$ 个观测值组成，所以总样本均值可利用式（13-5）求得。对于表13-1中的数据，我们得到下面的结果。

$$\bar{\bar{x}} = \frac{62 + 66 + 52}{3} = 60$$

因此，若原假设为真（$\mu_1 = \mu_2 = \mu_3 = \mu$），则总样本均值60为总体均值 μ 的最优估计值。

13.2.1 总体方差的处理间估计

在上一节，我们介绍了 σ^2 的一个处理间估计的概念，并且说明了当样本容量相等时如何计算处理间估计。

我们称 σ^2 的这个估计量为**均方处理**（mean square due to treatments，MSTR）。计算 MSTR 的一般公式为

$$\text{MSTR} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2}{k-1} \tag{13-6}$$

式（13-6）中的分子称为**处理平方和**（sum of squares due to treatments，SSTR）。分母 $k-1$ 表示与 SSTR 相联系的自由度。因此，均方处理也可以按以下公式计算。

均方处理

$$\text{MSTR} = \frac{\text{SSTR}}{k-1} \tag{13-7}$$

式中

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 \tag{13-8}$$

若 H_0 为真，则 MSTR 给出了 σ^2 的一个无偏估计。但是，如果 k 个总体均值不相等，则 MSTR 就不是 σ^2 的无偏估计；事实上，在这种情形下，MSTR 将会高估总体方差 σ^2。

对于表 13-1 中的 Chemitech 公司实验的数据，我们得到下面的结果。

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 5\times(62-60)^2 + 5\times(66-60)^2 + 5\times(52-60)^2 = 520$$

$$\text{MSTR} = \frac{\text{SSTR}}{k-1} = \frac{520}{2} = 260$$

13.2.2　总体方差的处理内估计

在上一节，我们介绍了 σ^2 的处理内估计的概念，并且说明了当样本容量相等时如何计算处理内估计。我们称 σ^2 的这个估计量为**均方误差**（mean square due to error，MSE）。计算 MSE 的一般公式为

$$\text{MSE} = \frac{\sum_{j=1}^{k} (n_j - 1)s_j^2}{n_T - k} \tag{13-9}$$

式（13-9）中的分子称为**误差平方和**（sum of squares due to error，SSE）。MSE 的分母是与 SSE 相联系的自由度。因此，MSE 的计算公式也可以表示成下面的形式。

均方误差

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} \tag{13-10}$$

式中

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 \tag{13-11}$$

我们注意到：MSE 是以每个处理内部的变异性为依据；它不受原假设是否为真的影响。因此，MSE 永远给出 σ^2 的一个无偏估计。

对于表 13-1 中 Chemitech 公司实验的数据，我们得到下面的结果。

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = (5-1)\times 27.5 + (5-1)\times 26.5 + (5-1)\times 31 = 340$$

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} = \frac{340}{15-3} = \frac{340}{12} = 28.33$$

13.2.3 方差估计量的比较：F 检验

如果原假设为真，则 MSTR 与 MSE 给出 σ^2 的两个独立的无偏估计量。根据在第 11 章中曾介绍过的内容，我们知道：对于正态总体，σ^2 的两个独立的估计量之比的抽样分布服从 F 分布。因此，如果原假设为真，并且 ANOVA 的假定得到满足，则 MSTR/MSE 的抽样分布服从一个分子自由度为 $k-1$，分母自由度为 n_T-k 的 F 分布。换句话说，如果原假设为真，则 MSTR/MSE 的值似乎应该从这个 F 分布中抽取。[⊖]

但是，如果原假设不成立，由于 MSTR 高估了总体方差 σ^2，从而使得 MSTR/MSE 的值被夸大。因此，如果得到的 MSTR/MSE 的值太大，以至于不像是随机抽取自分子自由度为 $k-1$，分母自由度为 n_T-k 的 F 分布的话，则我们将拒绝 H_0。因为，拒绝 H_0 的决定是基于 MSTR/MSE 的值，于是用来检验 k 个总体均值是否相等的检验统计量如下所示。

k 个总体均值相等的检验统计量

$$F = \frac{\text{MSTR}}{\text{MSE}} \tag{13-12}$$

检验统计量服从分子自由度为 $k-1$，分母自由度为 n_T-k 的 F 分布。

现在让我们回到 Chemitech 公司实验的例子，在 $\alpha=0.05$ 的显著性水平下进行假设检验。检验统计量的值为

$$F = \frac{\text{MSTR}}{\text{MSE}} = \frac{260}{28.33} = 9.18$$

分子自由度为 $k-1=3-1=2$，分母自由度为 $n_T-k=15-3=12$。因为对于大的检验统计量的值，我们将拒绝原假设，所以 p-值是检验统计量的值 $F=9.18$ 上侧的 F 分布曲线下方的面积。图 13-4 用图示说明了 $F=$ MSTR/MSE 的抽样分布，检验统计量的值，以及假设检验的 p-值，它是 F 分布上侧曲线下方的面积。[⊜]

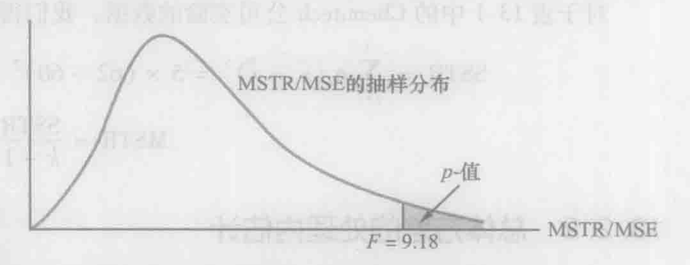


图 13-4 利用 MSTR/MSE 的抽样分布计算的 p-值

从附录 B 的表 A-4 中，我们可以找到如下表所示的分子自由度为 2，分母自由度为 12 的 F 分布上侧曲线下方面积所对应的 F 值。

上侧曲线下方的面积	0.10	0.05	0.025	0.01
F 值（$df_1=2$，$df_2=12$）	2.81	3.89	5.10	6.93

↑ $F=9.18$

因为 $F=9.18$ 大于 6.93，所以在 $F=9.18$ 处上侧曲线下方的面积小于 0.01。于是，p-值小于 0.01。Minitab 或 Excel 可以精确地计算出 p-值 $=0.004$。当 p-值 $\leqslant \alpha=0.05$ 时，则拒绝 H_0。检验提供了充分的证据使我们能够断定：三个总体均值是不相等的。换句话说，方差分析支持结论：三种装配方法每周生产的过滤系统总体的平均数量是不同的。

如同其他的检验假设方法一样，我们也可以利用临界值法。当 $\alpha=0.05$ 时，F 的临界值是分子自由度为 2，分母自由度为 12 的 F 分布上侧面积为 0.05 所对应的 F 值。从 F 分布表中，我们查出 $F_{0.05}=3.89$。于是，对 Chemitech 公司实验的例子，适用于上侧的拒绝法则是

$$\text{如果 } F \leqslant 3.89\text{，则拒绝 } H_0$$

因为 $F=9.18$，所以我们拒绝 H_0，并且得出结论：三个总体的均值是不相等的。检验 k 个总体均值相等的完

⊖ 我们已经在第 11.2 节中对 F 分布和 F 分布表的使用进行了介绍。

⊜ 附录 F 说明了如何使用 Minitab 或 Excel 计算 p-值。

整过程概括如下。

k个总体均值相等的检验

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a: k\text{个总体的均值不全相等}$$

检验统计量

$$F = \frac{\text{MSTR}}{\text{MSE}}$$

拒绝法则

$$p\text{-值法：如果}p\text{-值} \leqslant \alpha\text{，则拒绝}H_0$$
$$\text{临界值法：如果}F \geqslant F_\alpha\text{，则拒绝}H_0$$

式中，F_α 是分子自由度为 $k-1$，分母自由度为 n_T-k 时，使 F 分布的上侧曲线下方面积为 α 的 F 值。

13.2.4　ANOVA 表

前面的计算结果可以很方便地用方差分析表或 ANOVA 表（ANOVA table）表示出来。一个完全随机化实验设计的 ANOVA 表的一般形式如表 13-2 所示。表 13-3 是 Chemitech 公司实验的 ANOVA 表。在列标题“方差来源”中，与“总计”相联系的平方和被称为总平方和（SST）。我们注意到，Chemitech 公司实验的结果意味着：SST = SSTR + SSE，并且总平方和的自由度是处理平方和的自由度与误差平方和的自由度之和。

表 13-2　完全随机化设计的 ANOVA 表

方差来源	平方和	自由度	均方	F	p-值
处理	SSTR	$k-1$	$\text{MSTR} = \frac{\text{SSTR}}{k-1}$	$\frac{\text{MSTR}}{\text{MSE}}$	
误差	SSE	n_T-k	$\text{MSE} = \frac{\text{SSE}}{n_T-k}$		
总计	SST	n_T-1			

表 13-3　CHEMITECH 公司实验的方差分析表

方差来源	平方和	自由度	均方	F	p-值
处理	520	2	260.00	9.18	0.004
误差	340	12	28.33		
总计	860	14			

我们应该指出的是，如果我们将全部 15 个观测值看成是一个数据集，那么 SST 除以它的自由度 n_T-1，恰好是该数据集的总的样本方差。如果我们把整个数据集作为一个样本，总平方和 SST 的计算公式为

$$\text{SST} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{\bar{x}})^2 \tag{13-13}$$

可以证明，我们从 Chemitech 公司实验的方差分析表上看到的结果也可用于其他问题，即[⊖]

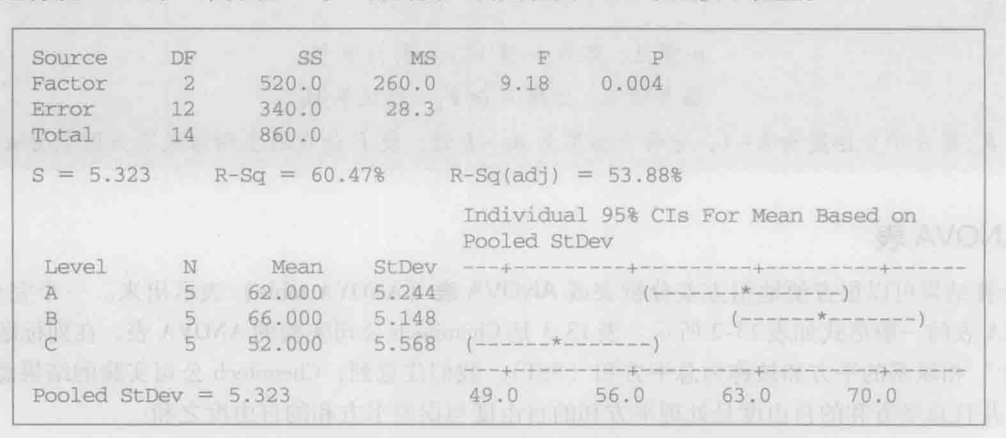
$$\text{SST} = \text{SSTR} + \text{SSE} \tag{13-14}$$

换句话说，SST 可以被分解为两个平方和：处理平方和与误差平方和。我们还注意到，SST 对应的自由度 n_T-1 也可以被分解为对应于 SSTR 的自由度 $k-1$ 与对应于 SSE 的自由度 n_T-k。方差分析可以被看作将总平方和及其自由度**分解**（partitioning）成它们对应的来源（处理与误差）的一个过程。这些平方和除以适当的自由度，可以给出方差的估计量，以及用于检验总体均值相等的假设的 F 值和 p-值。

⊖　方差分析可以被认为是将总平方和分解为不同成分的一种统计方法。

13.2.5 方差分析的计算机输出结果

利用计算机统计软件包，可以很容易地完成大样本容量或多个总体方差分析的计算。附录 13A、附录 13B 和附录 13C 分别给出了利用 Minitab、Excel 和 StatTools 实现方差分析计算的必要步骤。在图 13-5 中，我们给出了利用 Minitab 得到的 Chemitech 公司实验的输出结果。计算机输出的第一部分包含了我们熟悉的 ANOVA 表形式。比较一下图 13-5 与表 13-3，我们看到：尽管它们的标题略微有些不同，但可利用的信息却是相同的。图 13-5 的列标题“Source”对应于表 13-3 的“方差来源”的那一列，图 13-5 的行标题“Factor”则对应于表 13-3 的“处理”的那一行，并且将表 13-3 的“平方和”与“自由度”两列在图 13-5 中互换了位置。



```
Source      DF       SS       MS       F       P
Factor       2    520.0    260.0    9.18   0.004
Error       12    340.0     28.3
Total       14    860.0

S = 5.323     R-Sq = 60.47%     R-Sq(adj) = 53.88%

                                   Individual 95% CIs For Mean Based on
                                   Pooled StDev
 Level     N      Mean     StDev   ---+---------+---------+---------+------
 A         5    62.000     5.244            (-------*-------)
 B         5    66.000     5.148                  (-------*-------)
 C         5    52.000     5.568   (-------*-------)
                                   ---+---------+---------+---------+------
Pooled StDev = 5.323                49.0      56.0      63.0      70.0
```

图 13-5 CHEMITECH 公司实验的方差分析的 MINITAB 输出

注意：在图 13-5 中 ANOVA 表的下方，计算机的输出结果包含了各种装配方法的样本容量、样本均值和样本标准差。另外，Minitab 还给出了每个总体均值的 95% 的置信区间估计的图示。在建立这些置信区间估计时，Minitab 用 MSE 作为 σ^2 的估计。于是，MSE 的平方根给出了总体标准差 σ 的最佳估计。在计算机的输出中，σ 的这个估计量用“Pooled StDev”表示，它等于 5.323。为了详细解释这些区间估计是如何得来的，我们将以计算装配方法 A 的总体均值的 95% 的置信区间估计为例，加以说明。

我们从第 8 章区间估计的学习中知道，总体均值的区间估计的一般形式为

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \tag{13-15}$$

式中，s 为总体标准差 σ 的估计。因为 σ 的最佳估计是由“Pooled StDev”给出，所以我们将数值 5.323 代入式（13-15）中的 s。t 值的自由度是 12，它就是误差平方和的自由度。于是，$t_{0.025}=2.179$，我们得到

$$62 \pm 2.179 \times \frac{5.323}{\sqrt{5}} = 62 \pm 5.19$$

因此，装配方法 A 的置信水平为 95% 的置信区间是 62 − 5.19 = 56.81 至 62 + 5.19 = 67.19。因为 Chemitech 公司实验的例子中三种方法的样本容量都相等，所以装配方法 B 与装配方法 C 的置信区间也可由各自的样本均值加减 5.19 构成。因此，从 Minitab 所给出的输出图 13-5 中我们可以看到：这三个置信区间的宽度是相同的。

13.2.6 *k* 个总体均值相等的检验：一项观测性研究

我们已经说明了，对于一个完全随机化实验设计，怎样应用方差分析来检验 k 个总体均值是否相等的问题。重要的是要理解：ANOVA 也可以利用一项观测性研究得到的数据，来检验三个或三个以上总体均值是否相等的问题。作为一个例子，让我们考虑 National Computer Product（NCP）公司的情况。

NCP 公司在位于亚特兰大、达拉斯以及西雅图的工厂生产打印机与传真机。为了考察在这些工厂中有多少员工了解质量管理方面的知识，管理人员从每个工厂抽取一个由 6 名员工组成的随机样本，并对他们进行质量意识

考试。18 名员工的考试成绩列在表 13-4 中，每一组的样本均值、样本方差及样本标准差也在表中给出。管理人员希望利用这些数据来检验假设：三个工厂的平均考试成绩是相同的。

我们规定总体 1 为位于亚特兰大工厂的全体员工，总体 2 为位于达拉斯工厂的全体员工，总体 3 为位于西雅图工厂的全体员工。令

μ_1 = 总体 1 的平均考试成绩

μ_2 = 总体 2 的平均考试成绩

μ_3 = 总体 3 的平均考试成绩

尽管我们永远不可能知道 μ_1、μ_2、μ_3 的实际数值，但是我们仍然希望利用样本资料来检验下面的假设。

表 13-4　18 名员工的考试成绩

	工厂 1 亚特兰大	工厂 2 达拉斯	工厂 3 西雅图
	85	71	59
	75	75	64
	82	73	62
	76	74	69
	71	69	75
	85	82	67
样本均值	79	74	66
样本方差	34	20	32
样本标准差	5.83	4.47	5.66

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_a: \text{总体均值不全相等}$$

注意：对于 NCP 公司观测性研究进行的假设检验，与对于 Chemitech 公司的实验进行的假设检验是完全相同的。事实上，我们用于分析 Chemitech 公司实验的方差分析方法，也同样可以用来分析 NCP 公司的观测性研究得到的数据。⊖

即使在分析时使用同样的 ANOVA 方法，但值得注意的是，NCP 公司的观测性统计研究不同于 Chemitech 公司的实验性统计研究。进行 NCP 公司研究的人员无法控制如何将员工个人指派给工厂。也就是说，工厂已经处在正常的运转中，并且一名特定的员工已在三家工厂中的一家工作。NCP 公司所能做的就是从每一家工厂中抽取 6 名员工组成一个随机样本，并且实施质量意识考试。如果为了归类为实验性研究，NCP 公司就必须随机地抽取 18 名员工，然后以随机的方式将每一名员工指派到三家工厂中的一家工作。

注释和评论

1. 总样本均值也可以按照 k 个样本均值的加权平均值来计算。

$$\bar{\bar{x}} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_T}$$

在各个样本均值已给出的情况下，利用这个公式计算总样本均值较式（13-3）简单。

2. 如果每个样本都由 n 个观测值组成，则式（13-6）可以写作

$$\text{MSTR} = \frac{n\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{x}})^2}{k-1} = n\left[\frac{\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{x}})^2}{k-1}\right] = ns_{\bar{x}}^2$$

注意：当我们引入 σ^2 的处理间估计的概念时，这个结果与我们在 13.2 节中介绍的式（13-6）是相同的。式（13-6）是上面结果在样本容量不等时的简单推广。

3. 若每个样本有 n 个观测值，则 $n_T = kn$；于是，$n_T - k = k(n-1)$，并且式（13-9）可以重写成如下形式。

$$\text{MSE} = \frac{\sum_{j=1}^{k}(n-1)s_j^2}{k(n-1)} = \frac{(n-1)\sum_{j=1}^{k}s_j^2}{k(n-1)} = \frac{\sum_{j=1}^{k}s_j^2}{k}$$

换句话说，若样本容量相同，则 MSE 恰好为 k 个样本方差的算术平均值。注意：当我们引入 σ^2 的处理内估计的概念时，上式与我们在第 13.2 节中使用过的式（13-9）相同。

⊖ 练习第 8 题将要求您利用方差分析程序来分析 NCP 公司的数据。

练习

方法

2. 在一次完全随机化设计中，因子有5个水平，对于因子的每一个水平有7个实验单元。完成下面的ANOVA表。

方差来源	平方和	自由度	均方	F	p-值
处理	300				
误差					
总计	460				

4. 在一个检验三个不同处理的输出水平的实验设计中，得到了下列结果：SST = 400，SSTR = 150，$n_T = 19$。建立ANOVA表，并且在 $\alpha = 0.05$ 的显著性水平下，对三个处理的平均输出水平之间有无显著差异进行检验。

6. 对于下面的完全随机化实验设计进行方差分析计算。在 $\alpha = 0.05$ 的显著性水平下，处理均值之间是否存在显著差异？

	处理		
	A	B	C
	136	107	92
	120	114	82
	113	125	85
	107	104	101
	131	107	89
	114	109	117
	129	97	110
	102	114	120
		104	98
		89	106
$\bar{x}_j$	119	107	100
s_j^2	146.86	96.44	173.78

应用

8. 参阅表13-4的NCP公司的数据。建立ANOVA表。在 $\alpha = 0.05$ 的显著性水平下，检验三家工厂的平均考试成绩之间是否存在显著差异。

10. 审计人员必须以自己的直接经验、间接经验或者将两种经验的结合为依据，对一份审计报告的各个方面作出判断。在一项研究中，要求审计人员对一份审计报告中发现的错误频率作出判断。将审计人员作出的判断与实际结果进行比较。假设从一项类似的研究中得到下面的数据，较低的分数表示较好的判断力。

直接经验	间接经验	两种经验的结合
17.0	16.6	25.2
18.5	22.2	24.0
15.8	20.5	21.5
18.2	18.3	26.8
20.2	24.2	27.5
16.0	19.8	25.8
13.3	21.2	24.2

在 $\alpha = 0.05$ 的显著性水平下，检验判断的依据是否影响判断的质量。你有何结论？

12. 《消费者报告》杂志对提供全方位服务的餐厅连锁店进行了148 599次访问，完成了餐厅客户的满意度调查（*Consumer Reports* website）。在研究中，其中一个变量是吃一顿饭的价格，即平均每人吃一次晚餐及饮料支付的金额，不包含小费。假设，《太阳海岸时报》（*Sun Coast Times*）的一位记者认为，她的读者最感兴趣的问题是，对位于南卡罗来纳州默特尔海滩（Myrtle Beach）Grand Strand地区的餐厅进行一项类似的研究。该记者抽取了8家海鲜餐厅、8家意大利风味餐厅和8家牛排餐厅组成一个样本。对24家抽取出来的餐厅进行调查，得到吃一顿饭的价格数据如下表所示。在 $\alpha = 0.05$ 的显著性水平下，检验三种类型的餐厅，吃一顿饭的平均价格（单位：美元）之间是否存在显著差异。

意大利风味餐厅（美元）	海鲜餐厅（美元）	牛排餐厅（美元）
12	16	24
13	18	19
15	17	23
17	26	25
18	23	21
20	15	22
17	19	27
24	18	31

13.3 多重比较方法

当我们应用方差分析方法检验 k 个总体均值是否相等时，拒绝原假设只能让我们得出 k 个总体的均值不全相等的结论。在某些情况下，我们希望再向前迈进一步，并确定在 k 个均值中间到底哪几个均值之间存在差

异。本节的目的是说明，如何使用**多重比较方法**（multiple comparison procedures）在成对的总体均值之间进行统计比较。

13. 3. 1 Fisher 的 LSD 方法

假设，方差分析给出的统计依据，使得我们拒绝总体均值相等的原假设。在这种情况下，Fisher 的最小显著性差异（least significant difference，LSD）方法可以用来确定，到底在哪些均值之间存在着差异。为了说明在对总体均值进行两两比较时如何使用 Fisher 的 LSD 方法，我们仍然回到 13. 1 节介绍过的 Chemitech 公司实验的例子。利用方差分析，我们已经得出三种装配方法每周生产的过滤系统的平均数量是不相同的结论。在这种情况下，进一步的问题是：尽管我们相信三种装配方法每周生产的过滤系统的平均数量有差异，但差异出现在哪两种装配方法之间呢？也就是说，是总体 1 和总体 2 的均值不同，还是总体 1 和总体 3 的均值不同，抑或是总体 2 和总体 3 的均值不同呢？下面概况了成对总体均值进行比较的 Fisher 的 LSD 方法。

Fisher 的 LSD 方法

$$H_0: \mu_i = \mu_j$$
$$H_a: \mu_i \neq \mu_j$$

检验统计量

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\mathrm{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{13-16}$$

拒绝法则

p-值法：如果 p-值 $\leqslant \alpha$，则拒绝 H_0

临界值法：如果 $t \leqslant -t_{\alpha/2}$ 或者 $t \geqslant t_{\alpha/2}$，则拒绝 H_0

式中，$t_{\alpha/2}$是自由度为 $n_T - k$ 时，使 t 分布的上侧面积为 $\alpha/2$ 的 t 值。

现在，让我们在 $\alpha = 0.05$ 的显著性水平下，应用这一方法来判定总体 1（方法 A）和总体 2（方法 B）的均值之间是否存在显著的差异。由表 13-1 可知，方法 A 的样本均值为 62，方法 B 的样本均值为 66。由表 13-3 可知，MSE 的值为 28. 33，它是 σ^2 的估计，并且自由度为 12。对于 Chemitech 公司的实验数据，检验统计量的值是

$$t = \frac{62 - 66}{\sqrt{28.33\left(\frac{1}{5} + \frac{1}{5}\right)}} = -1.19$$

因为这是一个双边检验，p-值是 t 分布曲线下，$t = -1.19$ 下侧的面积的两倍。由附录 B 表 B-2 可知，自由度为 12 的 t 分布表给出了下面的信息。

上侧曲线下方面积	0. 20	0. 10	0. 05	0. 025	0. 01	0. 005
t 值（自由度 = 12）	0. 873	1. 356	1. 782	2. 179	2. 681	3. 055

↑
$t = 1.19$

t 分布表仅仅包含 t 的正值。但是，因为 t 分布是对称的，因此我们能够求出 t 分布曲线下，$t = 1.19$ 右侧的面积，两倍后就能得到对应于 $t = -1.19$ 的 p-值。我们看到，$t = 1.19$ 介于 0. 873 和 1. 356 之间，它们所对应的面积介于 0. 20 和 0. 10 之间，将对应的面积两倍后，得到 p-值介于 0. 40 和 0. 20 之间。Excel 或者 Minitab 都能给出精确的 p-值是 0. 257 1。因为 p-值大于 $\alpha = 0.05$，所以我们不能拒绝原假设。因此，我们不能得出结论：方法 A 每周生

产的过滤系统总体的平均数量与方法 B 每周生产的过滤系统总体的平均数量是不同的。[⊖]

许多有实际经验的专业人员发现，通过判断样本均值之差的大小而决定是否拒绝 H_0 更容易些。在这种情形下，检验统计量为 $\bar{x}_i - \bar{x}_j$，检验可按以下步骤进行。

基于检验统计量 $\bar{x}_i - \bar{x}_j$ 的 Fisher 的 LSD 方法

$$H_0: \mu_i = \mu_j$$

$$H_a: \mu_i \neq \mu_j$$

检验统计量

$$\bar{x}_i - \bar{x}_j$$

显著性水平 α 下的拒绝法则

$$\text{如果 } |\bar{x}_i - \bar{x}_j| > \text{LSD，则拒绝 } H_0$$

式中

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{13-17}$$

对于 Chemitech 公司实验的例子，LSD 的值为

$$\text{LSD} = 2.179\sqrt{28.33 \times \left(\frac{1}{5} + \frac{1}{5}\right)} = 7.34$$

注意：当样本容量相等时，只能计算出一个 LSD 的值。在这种情形下，我们可以简单地将任何两个样本均值之差的大小与 LSD 的值做比较。例如，总体 1（方法 A）与总体 3（方法 C）的样本均值之差为 62 - 52 = 10，因为该差值比 7.34 大，这就意味着，我们能够拒绝方法 A 与方法 C 每周生产的过滤系统总体的平均数量是相等的原假设。类似地，由于总体 2 与总体 3 的样本均值之差为 66 - 52 = 14 > 7.34，所以我们也能够拒绝方法 B 与方法 C 每周生产的过滤系统总体的平均数量是相等的原假设。实际上，我们的结论是：方法 A 和方法 B 这两种装配方法每周生产的过滤系统总体的平均数量与方法 C 每周生产的过滤系统总体的平均数量均不相同。

Fisher 的 LSD 方法也可用于建立两个总体均值之差的置信区间估计。一般步骤如下。

应用 Fisher 的 LSD 方法的两个总体均值之差的置信区间估计

$$\bar{x}_i - \bar{x}_j \pm \text{LSD} \tag{13-18}$$

式中

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{13-19}$$

其中，$t_{\alpha/2}$ 是自由度为 $n_T - k$ 时，使 t 分布的上侧面积为 $\alpha/2$ 的 t 值。

如果式（13-18）的置信区间包含数值 0，则我们不能拒绝两个总体均值相等的原假设。但是，如果式（13-18）的置信区间不包含数值 0，则我们可以得出两个总体均值之间存在差异的结论。对于 Chemitech 公司实验的例子，前面已经得到 LSD = 7.34（对应于 $t_{0.025} = 2.179$）。于是，总体 1 与总体 2 均值之差的置信水平为 95% 的置信区间估计是（62 - 66）± 7.34 = -4 ± 7.34 = -11.34 ~ 3.34；由于该区间包含数值 0，所以我们不能拒绝两个总体均值相等的假设。

⊖ 附录 F 给出了如何使用 Excel 或者 Minitab 计算 p-值的步骤。

13.3.2 第一类错误概率

我们开始讨论 Fisher 的 LSD 方法的前提是，方差分析为我们提供了拒绝总体均值相等的原假设的统计证据。在这种情形下，我们说明了如何使用 Fisher 的 LSD 方法来确定：总体均值之间的差异到底出现在哪些均值之间。在技术上，Fisher 的 LSD 方法被称为保护性或限制性 LSD 检验，这是因为，只有当我们首先找到一个用于方差分析的显著的 F 值时，我们才使用 LSD 检验。为了弄清楚在多重比较检验中，这种区别为什么是重要的，我们需要解释一下比较方式的第 I 类错误概率与实验方式的第 I 类错误概率的区别。

在 Chemitech 公司实验的例子中，我们利用 Fisher 的 LSD 方法做了三个成对的两两比较。

检验 1	检验 2	检验 3
$H_0: \mu_1 = \mu_2$	$H_0: \mu_1 = \mu_3$	$H_0: \mu_2 = \mu_3$
$H_a: \mu_1 \neq \mu_2$	$H_a: \mu_1 \neq \mu_3$	$H_a: \mu_2 \neq \mu_3$

在每一种情形下，我们都使用 $\alpha = 0.05$ 的显著性水平。因此，对每个检验，如果原假设为真，则犯第 I 类错误的概率为 $\alpha = 0.05$；因此，对每个检验，不犯第 I 类错误的概率就是 $1 - 0.05 = 0.95$。在讨论多重比较方法时，我们把这个第 I 类错误的概率（$\alpha = 0.05$）称为**比较方式的第 I 类错误概率**（comparisonwise Type I error rate）；比较方式的第 I 类错误概率表示了与单个的两两比较相联系的显著性水平。

现在我们考虑一个略为不同的问题。在进行三次成对的两两比较时，三次检验中至少有一次犯第I类错误的概率是多少？为回答这个问题，我们注意到：三次检验都不犯第I类错误的概率为 $0.95 \times 0.95 \times 0.95 = 0.8574$。[⊖]因此，至少有一次犯第 I 类错误的概率为 $1 - 0.8574 = 0.1426$。这样，当我们用 Fisher 的 LSD 方法进行三次成对的两两比较时，对应的犯第 I 类错误概率已经不是 0.05，其实是 0.1426；我们将这个错误概率称为总的或**实验方式的第 I 类错误概率**（experimentwise Type I error rate）。为避免混淆，我们将实验方式的第 I 类错误概率记为 α_{EW}。

对于总体个数较多的问题，犯实验方式第 I 类错误的概率就会变得比较大。例如，对于有 5 个总体，10 个可能成对的两两比较的问题。在比较方式的第 I 类错误概率为 $\alpha = 0.05$ 时，如果我们利用 Fisher 的 LSD 方法检验所有可能的成对的两两比较，则犯实验方式第 I 类错误的概率将是 $1 - (1 - 0.05)^{10} = 0.40$。在这种情形下，有实际经验的专业人员将会寻求其他方法以更好地控制犯实验方式第 I 类错误的概率。

控制总的犯实验方式错误概率的一种方法被称为 Bonferroni 修正方法，该方法在每一次检验中都使用一个较小的比较方式错误概率。例如，如果我们想要检验 C 个成对的两两比较，并希望总的犯实验方式第 I 类错误的最大概率为 α_{EW}，那么我们只要简单地将犯比较方式错误概率等于 α_{EW}/C 即可。在 Chemitech 公司实验的例子中，如果我们想要使用 Fisher 的 LSD 方法检验所有三个成对的两两比较，并且希望犯实验方式错误的最大概率为 $\alpha_{EW} = 0.05$ 时，那么我们只要设比较方式错误概率为 $\alpha = 0.05/3 = 0.017$ 即可。对于有 5 个总体，10 种可能的成对的两两比较的问题，Bonferroni 修正方法建议的比较方式错误概率为 $0.05/10 = 0.005$。回忆一下第 9 章有关假设检验的讨论，对于固定的样本容量，减少犯第 I 类错误的概率将导致增加犯第 II 类错误的概率，而第 II 类错误是指：当两个总体的均值实际上不相等时，却接受了两个总体均值相等的原假设。由于犯第 II 类错误的风险增加，因此，许多有实际经验的专业人员不愿在一个比较低的犯比较方式第 I 类错误的概率下进行单个的两两比较检验。

在这种情况下，还有几种其他方法，如 Tukey 方法与 Duncan 多重区域检验，也可用于解决这类问题。但是，究竟哪种方法是“最优的”，在统计学界存有较大争议。事实上，没有任何一种方法对所有类型的问题都是最优的。

⊖ 假设三次检验是独立的，因此我们可以简单地将各个概率连乘得到三个事件的联合概率。事实上，因为在每一次检验中都使用了均方误差 MSE，所以三次检验并不是独立的；因此，所涉及的犯错误的概率比已经表明的更大。

练习

方法

14. 下表的数据来自一个完全随机化设计。在下面的计算中，设 $\alpha=0.05$。

	处理 1	处理 2	处理 3
	63	82	69
	47	72	54
	54	88	61
	40	66	48
$\bar{x}_j$	51	77	58
s_j^2	96.67	97.34	81.99

a. 利用方差分析检验：3 个处理的均值之间是否存在显著的差异。

b. 利用 Fisher 的 LSD 方法确定哪些均值存在差异。

应用

☆16. 参阅第 15 题。利用 Fisher 的 LSD 方法，建立一个制造商 1 与制造商 2 的均值差的 95% 的置信区间估计。

18. 为检验 4 台机器发生故障的时间（单位：小时）之间是否存在显著差异，我们得到的数据如下表所示。

a. 在 $\alpha=0.05$ 的显著性水平下，4 台机器发生故障的总体平均时间之间，如果有差异的话，是什么样的差异？

b. 在 $\alpha=0.05$ 的显著性水平下，利用 Fisher 的 LSD 方法检验：机器 2 与机器 4 发生故障的总体平均时间是否相等。

机器 1	机器 2	机器 3	机器 4
6.4	8.7	11.1	9.9
7.8	7.4	10.3	12.8
5.3	9.4	9.7	12.1
7.4	10.1	10.3	10.8
8.4	9.2	9.2	11.3
7.3	9.8	8.8	11.5

20. 美国职业棒球小联盟有 14 支 3-A 级球队，这些球队分为北区、南区和西区。14 支球队在参加国际联盟（International League）的比赛时，每场比赛的平均观众人数如下表所示（The Biz of Baseball website，2009 年 1 月）。下表中还给出了球队的其他统计资料："W" 表示球队赢球，"L" 表示球队输球，"PCT" 表示球队赢球的比例。

a. 在 $\alpha=0.05$ 的显著性水平下，检验三个地区每场比赛的平均观众人数是否存在显著差异。

b. 在 $\alpha=0.05$ 的显著性水平下，利用 Fisher 的 LSD 方法，确定差异发生在哪些地区之间。

球队名称	地区	W	L	PCT	观众人数
Buffalo Bisons	北区	66	77	0.462	8 812
Lehigh Valley IronPigs	北区	55	89	0.382	8 479
Pawtucket Red Sox	北区	85	58	0.594	9 097
Rochester Red Wings	北区	74	70	0.514	6 913
Scranton-Wilkes Barre Yankees	北区	88	56	0.611	7 147
Syracuse Chiefs	北区	69	73	0.486	5 765
Charlotte Knights	南区	63	78	0.447	4 526
Durham Bulls	南区	74	70	0.514	6 995
Norfolk Tides	南区	64	78	0.451	6 286
Richmond Braves	南区	63	78	0.447	4 455
Columbus Clippers	西区	69	73	0.486	7 795
Indianapolis Indians	西区	68	76	0.472	8 538
Louisville Bats	西区	88	56	0.611	9 152
Toledo Mud Hens	西区	75	69	0.521	8 234

13.4　随机化区组设计

至此，我们已经讨论了完全随机化设计。回想一下，为检验处理均值之间的差异，我们计算 F 值时使用了比值

$$F = \frac{MSTR}{MSE} \tag{13-20}$$

每当外部因素（实验中没有考虑到的因素）引起的差异使得式（13-20）中的分母 MSE 变大时，将会出现一个问题。在这种情况下，式（13-20）的 F 值将会变小。于是，给我们发出的信号是处理均值之间不存在差异，而这样的差异事实上却是存在的。

本节我们将介绍一种被称为**随机化区组设计**（randomized block design）[㊀]的实验设计。这种实验设计方法是通过消除 MSE 项中来自外部的变异，以达到控制变异外部来源的目的。该设计方法为真实的误差方差给出了的一个更好的估计，并且在查明处理均值之间差异的能力方面，得到了一个更加有效的假设检验方法。作为阐述这种实验设计方法的例子，我们考虑空中交通管理员的工作压力问题。

13.4.1　空中交通管理员工作压力测试

一项测试空中交通管理员的疲劳程度与工作压力的研究得到的结果是建议改造并重新设计管理员工作站。考虑了工作站的若干设计方案后，三种最有可能减轻管理员工作压力的工作站具体方案被选出。关键问题是：三种方案对管理员工作压力的影响程度有多大差异？为了回答这个问题，我们需要设计一个实验，它能在每种工作站方案下给出空中交通管理员工作压力的测度。

在完全随机化实验中，管理员的随机样本被指派给每种工作站方案。但是，管理员们认为，在应对有压力的局面时，他们的能力是大不相同的。一名管理员认为是高压力而对于另一名管理员来说可能是中等压力或甚至是低压力。因此，当考虑变异的组内来源（MSE）时，我们必须意识到，该变异既包括随机误差又包括管理员个人差异导致的误差。事实上，管理者期望空中交通管理员个人的变异性是 MSE 项的一个主要贡献者。

将个人差异的影响分离出来的一种办法是使用随机化区组设计。[㊁]这样的设计能识别出源自管理员个人差异的变异性，并将其从 MSE 项中剔除。随机化区组需要管理员的一个单样本。样本中每个管理员要分别在三种工作站方案下接受检验。用实验设计的术语，工作站是影响因子，管理员是区组[㊂]。与工作站因子有关的三个处理或三个总体对应三种工作站方案。为简化起见，我们称工作站为系统 A、系统 B 和系统 C。

随机化区组设计中的随机化是指处理（系统）指派给管理员的顺序是随机的。如果每个管理员按照同样的顺序测试三个系统，任何观测到的系统间差异都可能归因于测试的顺序，而不是真正的系统差异。

为给出必要的数据，将三种工作站安装在俄亥俄州奥柏林的克利夫兰控制中心。随机选出 6 名管理员，并指派他们操作每个系统。对于参加研究的每一名管理员，随后要进行一次面试和一次医学体检，面试和体检提供了每名管理员在每个系统上工作压力的度量。数据如表 13-5 所示。

表 13-5　空中交通管理员工作压力测试的随机化区组设计

		处理		
		系统 A	系统 B	系统 C
区组	管理员 1	15	15	18
	管理员 2	14	14	14
	管理员 3	10	11	15
	管理员 4	13	12	17
	管理员 5	16	13	16
	管理员 6	13	13	13

表 13-6 是收集的管理员工作压力测试数据的汇总。在该表中，不但包括了列合计（处理）与行合计（区组），而且包括了在 ANOVA 方法中计算平方和时会用得到的一些样本均值。因为较低的工作压力测试值被认为是比较好，所以样本数据看起来似乎支持系统 B，因为它的平均压力等级为 13。但是，通常的问题依然存在：对于三个系统，样本数据证明了总体的平均压

㊀ 当实验单元的差异较小时，完全随机化设计是有用的。如果实验单元的差异较大时，通常利用**区组划分**（blocking）将实验单元划分成一些差异较小的组。

㊁ 商务中的实验性研究通常包含具有高度差异性的实验单元，因此，经常使用随机化区组设计。

㊂ 实验设计中的区组划分与抽样中的分层类似。

力水平是不同的结论吗？也就是说，总体平均压力水平的差异在统计上显著吗？类似完全随机化设计方法，方差分析的计算可以用来回答这个统计问题。

表 13-6 空中交通管理员工作压力测试的压力数据汇总

		处理			行或区组合计	区组均值
		系统 A	系统 B	系统 C		
区组	管理员 1	15	15	18	48	$\bar{x}_{1.}=48/3=16.0$
	管理员 2	14	14	14	42	$\bar{x}_{2.}=42/3=14.0$
	管理员 3	10	11	15	36	$\bar{x}_{3.}=36/3=12.0$
	管理员 4	13	12	17	42	$\bar{x}_{4.}=42/3=14.0$
	管理员 5	16	13	16	45	$\bar{x}_{5.}=45/3=15.0$
	管理员 6	13	13	13	39	$\bar{x}_{6.}=39/3=13.0$
列或处理合计		81	78	93	252	$\bar{\bar{x}}=\frac{252}{18}=14.0$
处理均值		$\bar{x}_{.1}=\frac{81}{6}=13.5$	$\bar{x}_{.2}=\frac{78}{6}=13.0$	$\bar{x}_{.3}=\frac{93}{6}=15.5$		

13.4.2 ANOVA 方法

随机化区组设计的 ANOVA 方法要求我们将总平方和（SST）分解成三个部分：处理平方和（SSTR），区组平方和（SSBL）和误差平方和（SSE）。该分解的公式如下。

$$SST = SSTR + SSBL + SSE \tag{13-21}$$

我们将这一平方和的分解汇总于表 13-7 所示的随机化区组设计的 ANOVA 表中。表中所使用的记号分别表示：k 代表处理的个数；b 代表区组的个数；n_T 代表总样本容量（$n_T=kb$）。

表 13-7 k 个处理，b 个区组的随机化区组设计的 ANOVA 表

方差来源	平方和	自由度	均方	F	p-值
处理	SSTR	$k-1$	$MSTR=\frac{SSTR}{k-1}$	$\frac{MSTR}{MSE}$	
区组	SSBL	$b-1$	$MSBL=\frac{SSBL}{b-1}$		
误差	SSE	$(k-1)(b-1)$	$MSE=\frac{SSE}{(k-1)(b-1)}$		
总计	SST	n_T-1			

注意：ANOVA 表还显示出，总自由度 n_T-1 如何被分解成处理的自由度 $k-1$、区组的自由度 $b-1$ 和误差项的自由度 $(k-1)(b-1)$ 的和。均方列表示的是平方和被自由度除，$F=MSTR/MSE$ 是用于检验处理均值之间差异显著性的 F 比。随机化区组设计的主要贡献是：通过划分区组，我们将管理员个人的差异从 MSE 项中剔除，并且得到了一个在三种可供选择的工作站方案中压力差异的更有力的检验。

13.4.3 计算与结论

为了检验随机化区组设计中的处理均值之间差异，我们需要计算 F 统计量。为了计算 F 统计量，需要计算 MSTR 和 MSE。为了计算这两个均方值，我们必须首先计算 SSTR 和 SSE；此外，我们还要计算 SSBL 和 SST。为简化演示，我们分四步完成计算。除了 k，b，n_T 定义如前外，另外还使用了下列记号：

x_{ij} 代表在区组 i 中对应于处理 j 的观测值；$\bar{x}_{.j}$ 代表第 j 个处理的样本均值；$\bar{x}_{i.}$ 代表第 i 个区组的样本均值；$\bar{\bar{x}}$ 代表总样本均值。

第 1 步：计算总平方和（SST）。

$$SST = \sum_{i=1}^{b}\sum_{j=1}^{k}(x_{ij} - \bar{\bar{x}})^2 \tag{13-22}$$

第 2 步：计算处理平方和（SSTR）。

$$SSTR = b\sum_{j=1}^{k}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13-23}$$

第 3 步：计算区组平方和（SSBL）。

$$SSBL = k\sum_{i=1}^{b}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13-24}$$

第 4 步：计算误差平方和（SSE）。

$$SSE = SST - SSTR - SSBL \tag{13-25}$$

对于表 13-6 中的空中交通管理员的数据，根据这些计算步骤得到下面的平方和。

第 1 步：$SST = (15-14)^2 + (15-14)^2 + (18-14)^2 + \cdots + (13-14)^2 = 70$

第 2 步：$SSTR = 6 \times [(13.5-14)^2 + (13.0-14)^2 + (15.5-14)^2] = 21$

第 3 步：$SSBL = 3 \times [(16-14)^2 + (14-14)^2 + (12-14)^2 + (14-14)^2 + (15-14)^2 + (13-14)^2] = 30$

第 4 步：$SSE = 70 - 21 - 30 = 19$

这些平方和分别除以它们各自的自由度，得到如表 13-8 所示的对应的均方值。

表 13-8　空中交通管理员工作压力测试的 ANOVA 表

方差来源	平方和	自由度	均方	*F*	*p*-值
处理	21	2	10.5	10.5/1.9 = 5.53	0.024
区组	30	5	6.0		
误差	19	10	1.9		
总计	70	17			

在 $\alpha = 0.05$ 的显著性水平下，我们进行假设检验。检验统计量的值是

$$F = \frac{MSTR}{MSE} = \frac{10.5}{1.9} = 5.53$$

检验统计量分子的自由度是 $k-1=3-1=2$，分母的自由度是 $(k-1)(b-1)=(3-1)(6-1)=10$。因为我们仅是对大的检验统计量的值拒绝原假设，*p*-值是对应于 $F=5.53$ 的 *F* 分布上侧曲线下方的面积。从附录 B 的表 B-4我们查出：对于分子自由度为 2 和分母自由度为 10 的 $F=5.53$ 介于 $F_{0.025}=5.46$ 和 $F_{0.01}=7.56$ 之间。于是，我们得知：*F* 分布上侧的面积或 *p*-值介于 0.01 和 0.025 之间。我们可以使用 Excel 或 Minitab 得到对应于 $F=5.53$ 的精确的 *p*-值是 0.024。由于 *p*-值 $\leqslant \alpha = 0.05$，所以我们拒绝原假设 $H_0: \mu_1 = \mu_2 = \mu_{\mu 3}$，并且得出结论：对于三个可供选择的工作站，总体的平均压力水平是不同的。

关于随机化区组设计，我们可以做一些一般性的注释。本节所介绍的实验设计是一种完全的区组设计；“完全”一词表明：每一个区组都要从属于所有 *k* 个处理。就是说，所有管理员（区组）要在所有三个系统（处理）下接受测试。如果某些（但不是全部）处理被用于每个区组，这样的实验设计被称为不完全区组设计。关于不完全区组设计的讨论已经超出了本书的范围。

因为在空中交通管理员工作压力的测试中，我们要求每一名管理员要使用所有三个系统，这一要求保证了该设计为完全区组设计。但是，在某些情况下，区组的划分是在每个区组内使实验单元是“相似的”情况下而实施的。例如，假设在空中交通管理员的一个预先测试中，管理员总体被分为若干个组，从极高的个人压力组到极低的个人压力组。在每一个压力分组中有三名管理员参加研究，这样我们就实现了区组划分。这样每个区组将由同一压力组的三名管理员组成。区组设计中的随机化就变成将每一区组中的三名管理员随机地指派给三个系统。

最后，我们要注意：表 13-7 所示的 ANOVA 表给出的 *F* 值是为了检验处理的影响，而不是检验区组的影响。其原因在于实验是为了检验单一因素——工作站方案而设计的。基于个人压力差异的区组划分是为了从 MSE 项中剔除这种变异性而实施的。但是，该研究不是特别地为检验管理员个人压力差异而设计的。

有些统计分析员计算 F = MSBL/MSE，并用该统计量检验区组的显著性。然后他们利用这一结果作为在将来的实验中是否可以要求进行同样类型的区组划分的一个根据。然而，如果管理员个人压力差异成为研究的一个因子，那么应该使用不同的实验设计。有关第二个因子的结论，不应该作为实施区组显著性检验的一个依据。

注释和评论

因为有 b 个区组使得自由度减少了 $b-1$，所以随机化区组设计的误差自由度小于完全随机化设计的误差自由度。如果 n 很小，因为误差自由度的减少，区组的潜在影响可能被掩盖；当 n 很大时，这种影响被最小化了。

练习

方法

22. 若一个随机化区组设计包含 5 个处理和 3 个区组，经计算，得到如下所示的数据：SST = 430，SSTR = 310，SSBL = 85。建立 ANOVA 表，并在 $\alpha = 0.05$ 的显著性水平下，检验是否存在显著性差异。

应用

24. 一位汽车经销商对调整一台小型发动机所需时间（单位：分）进行了测试，以判断所需的调整时间是否与使用计算机控制的发动机分析器还是电子控制的发动机分析器有关。因为小型、中型和大型汽车发动机的调整时间是不同的，因此我们将三种类型的汽车作为实验的区组。得到的数据如下表所示。

		分析器	
		计算机控制的	电子控制的
汽车	小型	50	42
	中型	55	44
	大型	63	46

在 $\alpha = 0.05$ 的显著性水平下，检验两种分析器所需的调整时间是否存在显著性差异。

26. 学习能力倾向测验（SAT）包括三个部分：评论性阅读、数学和写作。每一部分的满分为 800 分。有关 2009 版的 SAT 测试分数的统计资料可以从美国大学理事会的网站（College Board website）上得到。由 6 名学生的 SAT 测试分数组成的样本数据如下表所示。

学生	评论性阅读	数学	写作
1	526	534	530
2	594	590	586
3	465	464	445
4	561	566	553
5	436	478	430
6	430	458	420

a. 在 $\alpha = 0.05$ 的显著性水平下，6 名学生的 SAT 测试成绩存在差异吗？

b. 哪一部分的测试似乎给学生们带来最大忧虑？请作出解释。

13.5 析因实验

迄今为止，我们所讨论的实验设计使我们能够得出有关一个因子的一些统计结论。然而，在有些实验中，我们希望得到有关一个以上变量或因子的统计结论。**析因实验**（factorial experiment）是一种实验设计，该实验设计允许我们同时得到有关两个或两个以上因子同时存在时的一些统计结论。之所以使用术语“析因”是因为实验条件包括了所有可能的因子组合。例如，如果有因子 A 的 a 个水平，因子 B 的 b 个水平，那么实验将涉及收集 ab 个处理组合的数据。本节我们将介绍两因子析因实验的分析。其基本方法可以推广到两个以上因子的析因实验。

作为两因子析因实验的一个例子，我们将考虑与管理类研究生入学考试（Graduate Management Admissions Test，GMAT）有关的一项研究。GMAT 是一种商学院的研究生院用来评价攻读该领域研究生课程的考生能力的标准化考试，其分数在 200 ~ 800 之间，分数越高表明能力越强。

在试图提高学生的 GMAT 的成绩时，著名的得克萨斯大学考虑提供下面三种 GMAT 辅导课程。

1. 3 小时的复习课，内容覆盖了 GMAT 常考的题型。
2. 一天的课程，内容覆盖了相关的考试内容，还要进行一次模拟考试并评定分数。

3. 10 周的强化班，内容涉及发现每个考生的薄弱环节并设立个性化的提高课程。

因此，这项研究中的一个因子就是 GMAT 辅导课程，该课程有三个处理：3 小时的复习课、一天的课程和 10 周的强化班。在选择采取哪种辅导课程之前，将进行进一步的研究以判断这三种被推荐的辅导课程是如何影响 GMAT 分数的。

通常参加 GMAT 考试的学生来自三种类型的院校：商学院、工学院以及艺术与科学学院。因此，在实验中值得关注的第二个因子就是学生本科所在的院校是否影响 GMAT 分数。本科所在院校这第二个因子，也有三个处理：商学院、工学院以及艺术与科学学院。对于这一实验的析因设计有三个处理对应于因子 A——辅导课程，有三个处理对应于因子 B——本科所在的院校，因此一共有 3 × 3 = 9 种处理组合。这些处理组合或实验条件汇总于表 13-9 中。

假设，对应于表 13-9 所示的 9 种处理组合的每一个，我们都选取两名学生组成一个样本：两名商学院学生参加 3 小时复习课、另两名参加一天的课程，还有两名参加 10 周强化班。此外，对于每种辅导课程，各有两名工学院的学生以及艺术与科学学院的学生参加。用实验设计的术语，每个处理组合容量为 2 的样本意味着我们有两个**复制**（replications）。使用更多数量的复制和更大容量的样本是件很容易的事情，但是，为了说明这一方法，我们作出了使本例计算极小化的选择，因此只有两个复制。

表 13-9 两因子 GMAT 实验的 9 种处理组合

		因子 B：本科院校		
		商学院	工学院	艺术与科学学院
因子 A：辅导课程	3 小时复习课	1	2	3
	1 天的课程	4	5	6
	10 周强化班	7	8	9

表 13-10 两因子实验的 GMAT 分数

		因子 B：本科院校		
		商学院	工学院	艺术与科学学院
因子 A：辅导课程	3 小时复习班	500	540	480
		580	460	400
	1 天的课程	460	560	420
		540	620	480
	10 周强化班	560	600	480
		600	580	410

这个实验设计要求：从每一个本科院校准备读研究生的学生里随机地抽取 6 名学生。然后，每个本科院校的两名学生应该被随机地指派参加每种辅导课程，从而一共有 18 名学生参加了该项研究。

我们假定：随机抽取的学生已经上完辅导课程，并且也已参加了 GMAT。他们的得分列于表 13-10 中。

我们对表 13-10 中的数据进行方差分析计算，计算结果可以给出下面问题的答案。

- 主影响（因子 A）：辅导课程的不同对 GMAT 的成绩有影响吗？
- 主影响（因子 B）：本科院校的不同对 GMAT 的成绩有影响吗？
- 交互影响（因子 A 和 B）：一些本科院校的学生参加某种类型的辅导课程得到较好的 GMAT 成绩，而另一些本科院校的学生参加另一种类型的辅导课程能得到较好的 GMAT 成绩吗？

术语交互作用（interaction）是指一种新的影响，因为我们应用了析因实验，所以现在可以研究这种新的影响了。如果交互作用的影响对于 GMAT 分数有显著作用，那么我们可以得到结论：辅导课程类型的影响取决于本科院校。

13.5.1 ANOVA 方法

两因子析因实验的 ANOVA 方法要求我们将总平方和（SST）分解为四个部分：因子 A 的平方和（SSA），因子 B 的平方和（SSB），交互作用的平方和（SSAB）和误差平方和（SSE）。分解公式如下。

$$SST = SSA + SSB + SSAB + SSE \tag{13-26}$$

平方和与自由度的分解汇总于表13-11 中，表中使用了下面的符号：a 代表因子 A 的水平数；b 代表因子 B 的水平数；r 代表复制的个数；n_T 代表实验中观测值的总数；$n_T = abr$。

表 13-11 有 r 个复制的两因子析因实验的 ANOVA 表

方差来源	平方和	自由度	均方	F	p-值
因子 A	SSA	$a-1$	$MSA=\frac{SSA}{a-1}$	$\frac{MSA}{MSE}$	
因子 B	SSB	$b-1$	$MSB=\frac{SSB}{b-1}$	$\frac{MSB}{MSE}$	
交互作用	SSAB	$(a-1)(b-1)$	$MSAB=\frac{SSAB}{(a-1)(b-1)}$	$\frac{MSAB}{MSE}$	
误差	SSE	$ab(r-1)$	$MSE=\frac{SSE}{ab(r-1)}$		
总计	SST	n_T-1			

13.5.2 计算与结论

为了检验因子 A、因子 B 和交互作用的显著性，我们需要计算 F 统计量，为此我们需要计算 MSA、MSB、MSAB 和 MSE。为计算这四个均方，我们必须首先计算 SSA、SSB、SSAB 和 SSE；与此同时，我们还要计算 SST。为了简化表述，我们将计算分 5 步进行。除 a、b、r 和 n_T 如前定义外，我们还使用了下面的符号。

x_{ijk}——对应于因子 A 的处理 i 和因子 B 的处理 j 的第 k 次复制的观测值

$\bar{x}_{i.}$——处理 i（因子 A）的观测值的样本均值

$\bar{x}_{.j}$——处理 j（因子 B）的观测值的样本均值

$\bar{x}_{ij}$——对应于处理 i（因子 A）和处理 j（因子 B 的）组合的观测值的样本均值

$\bar{\bar{x}}$——所有 n_T 个观测值的总样本均值

第 1 步：计算总平方和。

$$SST=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk}-\bar{\bar{x}})^2 \tag{13-27}$$

第 2 步：计算因子 A 的平方和。

$$SSA=br\sum_{i=1}^{a}(\bar{x}_{i.}-\bar{\bar{x}})^2 \tag{13-28}$$

第 3 步：计算因子 B 的平方和。

$$SSB=ar\sum_{j=1}^{b}(\bar{x}_{.j}-\bar{\bar{x}})^2 \tag{13-29}$$

第 4 步：计算交互作用的平方和。

$$SSAB=r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}})^2 \tag{13-30}$$

第 5 步：计算误差平方和。

$$SSE=SST-SSA-SSB-SSAB \tag{13-31}$$

表 13-12 列出了实验所收集的数据和有助于我们计算平方和的各个合计。对于 GMAT 两因子析因实验，我们利用式（13-27）~式（13-31），计算的平方和如下所示。

第 1 步：$SST=(500-515)^2+(580-515)^2+(540-515)^2+\cdots+(410-515)^2=82\,450$

第 2 步：$SSA=3\times2\times[(493.33-515)^2+(513.33-515)^2+(538.33-515)^2]=6\,100$

第 3 步：$SSB=3\times2\times[(540-515)^2+(560-515)^2+(445-515)^2]=45\,300$

第 4 步：$SSAB=2\times[(540-493.33-540+515)^2+(500-493.33-560+515)^2+\cdots+(445-538.33-445+515)^2]=11\,200$

第 5 步：$SSE=82\,450-6\,100-45\,300-11\,200=19\,850$

表 13-12 两因子实验的 GMAT 汇总数据

处理组合合计		因子 B：本科院校			行合计	因子 A 平均值
		商学院	工学院	艺术与科学学院		
因子 A：辅导课程	3 小时复习课	500 580 1 080 $\bar{x}_{11}=\frac{1\,080}{2}=540$	540 460 1 000 $\bar{x}_{12}=\frac{1\,000}{2}=500$	480 400 880 $\bar{x}_{13}=\frac{880}{2}=440$	2 960	$\bar{x}_{1}.=\frac{2\,960}{6}=493.33$
	1 天的课程	460 540 1 000 $\bar{x}_{21}=\frac{1\,000}{2}=500$	560 620 1 180 $\bar{x}_{22}=\frac{1\,180}{2}=590$	420 480 900 $\bar{x}_{23}=\frac{900}{2}=450$	3 080	$\bar{x}_{2}.=\frac{3\,080}{6}=513.33$
	10 周强化班	560 600 1 160 $\bar{x}_{31}=\frac{1\,160}{2}=580$	600 580 1 180 $\bar{x}_{32}=\frac{1\,180}{2}=590$	480 410 890 $\bar{x}_{33}=\frac{890}{2}=445$	3 230	$\bar{x}_{3}.=\frac{3\,230}{6}=538.33$
列项合计		3 240	3 360	2 670	9 270 ← 合计	
因子 B 平均值		$\bar{x}_{.1}=\frac{3\,240}{6}=540$	$\bar{x}_{.2}=\frac{3\,360}{6}=560$	$\bar{x}_{.3}=\frac{2\,670}{6}=445$	$\bar{\bar{x}}=\frac{9\,270}{18}=515$	

这些平方和它们所对应的自由度相除，得到适用于检验两个主影响（辅导课程和本科院校）和交互作用影响的均方值。

因为在任何中型到大型的析因实验中涉及大量的计算，因此在实施上面方差分析的计算和用于假设检验推断的 p-值的计算时，计算机通常起着重要的作用。图 13-6 显示了 GMAT 两因子析因实验方差分析的 Minitab 输出结果。让我们利用 Minitab 的输出结果，在 $\alpha=0.05$ 的显著性水平下，对两因子的 GMAT 研究进行假设检验。用于检验三种辅导课程（因子 A）之间显著差异的 p-值是 0.299。因为 p-值 $=0.299$，大于 $\alpha=0.05$，所以对于三种辅导课程，GMAT 的平均考试成绩不存在显著差异。但是，对于本科院校的影响，p-值 $=0.005$，小于 $\alpha=0.05$；于是，对于三种类型的本科院校，GMAT 的平均考试成绩存在显著差异。

```
SOURCE         DF      SS      MS       F       P
Factor A        2    6100    3050    1.38   0.299
Factor B        2   45300   22650   10.27   0.005
Interaction     4   11200    2800    1.27   0.350
Error           9   19850    2206
Total          17   82450
```

图 13-6 GMAT 两因子析因实验设计的 MINITAB 输出

最后，因为交互作用影响的 p-值 $=0.350$，大于 $\alpha=0.05$，所以不存在显著的交互作用影响。综上所述，这项研究没有理由让我们相信：对于来自不同本科院校准备参加 GMAT 考试的学生，三种辅导课程在提高他们 GMAT 的成绩方面是不同的。

我们发现，本科院校是一个显著的因子。查看表 13-12 中的计算，我们看到样本均值是：商学院学生 $\bar{x}_{.1}=540$，工学院学生 $\bar{x}_{.2}=560$，艺术与科学学院学生 $\bar{x}_{.3}=445$。我们可以对个别处理的均值进行检验；然而，观测三个样本均值后，我们可以期望：商学院和工学院学生的 GMAT 分数对辅导课程没有显著区别。但是，艺术与科学学院学生的 GMAT 分数看起来要显著地低于其他院校的学生。也许这种看法会促使艺术与科学学院类的大学考虑其他的途径，以帮助这些学生准备 GMAT。

练 习

方法

☆28. 一项涉及因子 A 的两个水平与因子 B 的三个水平的析因实验得到下面的数据。

		因子 B		
		水平 1	水平 2	水平 3
因子 A	水平 1	135	90	75
		165	66	93
	水平 2	125	127	120
		95	105	136

在 $\alpha=0.05$ 的显著性水平下，检验主影响与交互作用影响的显著性。

应用

30. 一家经营邮购商品广告目录的公司设计了一个析因实验，以检验在杂志上刊登广告的尺寸大小和广告的设计图案，对于要求得到邮购商品目录的数量（单位：1 000 份）的影响。考虑三种广告的设计图案和两种不同的广告尺寸，得到以下数据。利用析因设计的 ANOVA 方法，在 $\alpha=0.05$ 的显著性水平下，检验广告的设计图案、广告的尺寸，以及交互作用的显著性影响。

		广告尺寸	
		小	大
设计图案	A	8	12
		12	8
	B	22	26
		14	30
	C	10	18
		18	14

32. 作为 *Consumer Reports* 杂志设计研究的一部分是要对混合动力汽车和具有同样配置的传统动力汽车进行比较。于是，该杂志对各种型号的混合动力、传统动力的汽车和 SUV 进行了测试。下面是 *Consumer Reports* 经过测试取得的 16 辆各种型号汽车的油耗等级（单位：英里/加仑）的数据，包括两辆混合动力小型汽车、两辆混合动力中型汽车、两辆混合动力小型 SUV、两辆混合动力中型 SUV；以及 8 辆具有同样配置的传统动力汽车（*Consumer Reports*, 2008 年 10 月）。

品牌/型号	车型	动力类型	油耗（英里/加仑）
Honda Civic	小型车	混合动力	37
Honda Civic	小型车	传统动力	28
Toyota Prius	小型车	混合动力	44
Toyota Corolla	小型车	传统动力	32
Chevrolet Malibu	中型车	混合动力	27
Chevrolet Malibu	中型车	传统动力	23
Nissan Altima	中型车	混合动力	32
Nissan Altima	中型车	传统动力	25
Ford Escape	小型 SUV	混合动力	27
Ford Escape	小型 SUV	传统动力	21
Sature Vue	小型 SUV	混合动力	28
Sature Vue	小型 SUV	传统动力	22
Lexus RX	中型 SUV	混合动力	23
Lexus RX	中型 SUV	传统动力	19
Toyota Highlander	中型 SUV	混合动力	24
Toyota Highlander	中型 SUV	传统动力	18

在 $\alpha=0.05$ 的显著性水平下，检验车型、动力类型以及它们之间交互作用的显著性影响。

小 结

本章我们介绍了如何利用方差分析来检验若干个总体或处理之间的差异。我们引进了完全随机化设计、随机化区组设计和两因子析因实验。我们利用完全随机化设计和随机化区组设计，可以得出有关单个因子均值之间是否存在差异的结论。在随机化区组设计中划分区组的主要目的是从误差项中剔除来自外部的变异。这样的区组划分给出了真实误差方差一个更好的估计，并且给出了确定该因子的总体或处理之间的均值是否存在显著差异的一个更好的检验。

我们已经说明了，在方差分析和实验设计中使用的统计检验的基础是建立总体方差 σ^2 的两个独立的估计量。在单因子情形，一个估计量以处理之间的差异为依据；这个估计量仅当均值 μ_1, μ_2, …, μ_k 全相等时，给出 σ^2 的一个无偏估计量。σ^2 的第二个估计量以每个样本内部观测值的差异为依据；这个估计量总能给出 σ^2 的一个无偏估计量。为了确定总体或处理均值相等的原假设是否成立，我们通过计算 σ^2 的这两个估计量的比值（F 统计量），建

立了一个拒绝该原假设的规则。在所有讨论过的实验设计中，我们要按照不同的来源对平方和与自由度进行分解，从而使我们能够计算出适用于方差分析和检验的数值。我们还说明了 Fisher 的 LSD 方法和 Bonferroni 修正方法如何用于成对比较以判断在哪些均值之间存有差异。

关键术语

factor 因子 引起关注的自变量的另一个称谓。

treatments 处理 因子的不同水平。

single-factor experiment 单因子实验 只涉及有 k 个总体或处理的一个因子的实验。

response variable 响应变量 引起关注的应变量的另一个称谓。

experimental units 实验单元 实验中引起关注的研究对象。

completely randomized design 完全随机化设计 处理被随机地指派给实验单元的一种实验设计。

ANOVA table ANOVA 表 一种用来汇总方差分析计算和结果的表。它包括显示方差来源、平方和、自由度、均方和 F 值的列。

partitioning 分解 将总平方和与自由度分配给各组成部分的过程。

multiple comparison procedures 多重比较方法 能用于成对的总体均值之间进行统计比较的统计方法。

comparisonwise Type Ⅰ error rate 比较方式的第Ⅰ类错误概率 与单个两两成对比较相联系的犯第Ⅰ类错误的概率。

experimentwise Type Ⅰ error rate 实验方式的第Ⅰ类错误概率 若干个两两比较中至少有一个犯第Ⅰ类错误的概率。

blocking 区组划分 对所有的处理使用相同的或相似的实验单元的过程。区组划分的目的是从误差项中删除来自外部的变异，因此给出了总体或处理均值之间是否存在差异的更有力的检验。

randomized block design 随机化区组设计 使用区组划分的一种实验设计

factorial experiment 析因实验 一种实验设计方法，该方法允许我们得到有关两个或两个以上因子同时存在的统计结论。

replications 复制 在一个实验中每个实验条件的重复次数。

interaction 交互作用 当一个因子的水平与另一个因子的水平相互作用时，对响应变量产生的影响。

重要公式

完全随机化设计

处理 j 的样本均值

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \tag{13-1}$$

处理 j 的样本方差

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1} \tag{13-2}$$

总样本均值

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T} \tag{13-3}$$

$$n_T = n_1 + n_2 + \cdots + n_k \tag{13-4}$$

均方处理

$$\text{MSTR} = \frac{\text{SSTR}}{k-1} \tag{13-7}$$

处理平方和

$$\text{SSTR} = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2 \tag{13-8}$$

均方误差

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} \tag{13-10}$$

误差平方和

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1) s_j^2 \tag{13-11}$$

k 个总体均值相等的检验统计量

$$F = \frac{\text{MSTR}}{\text{MSE}} \tag{13-12}$$

总平方和

$$\text{SST} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{\bar{x}})^2 \tag{13-13}$$

平方和分解

$$\text{SST} = \text{SSTR} + \text{SSE} \tag{13-14}$$

多重比较方法

Fisher 的 LSD 方法的检验统计量

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{13-16}$$

Fisher 的 LSD

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{13-17}$$

随机化区组设计

总平方和

$$\text{SST} = \sum_{i=1}^{b}\sum_{j=1}^{k}(x_{ij} - \bar{\bar{x}})^2 \tag{13-22}$$

处理平方和

$$\text{SSTR} = b\sum_{j=1}^{k}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13-23}$$

区组平方和

$$\text{SSBL} = k\sum_{i=1}^{b}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13-24}$$

误差平方和

$$\text{SSE} = \text{SST} - \text{SSTR} - \text{SSBL} \tag{13-25}$$

析因实验

总平方和

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk} - \bar{\bar{x}})^2 \tag{13-27}$$

因子 A 的平方和

$$\text{SSA} = br\sum_{i=1}^{a}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13-28}$$

因子 B 的平方和

$$\text{SSB} = ar\sum_{j=1}^{b}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13-29}$$

交互作用平方和

$$\text{SSAB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2 \tag{13-30}$$

误差平方和

$$\text{SSE} = \text{SST} - \text{SSA} - \text{SSB} - \text{SSAB} \tag{13-31}$$

补充练习

34. 在一项完全随机化实验设计中，检验三种品牌的纸巾它们的吸水能力。我们对相同规格的纸巾，每种品牌选取四小块进行检验。吸水性的等级数据如下表所示。在 $\alpha = 0.05$ 的显著性水平下，三种品牌纸巾的吸水能力存在差异吗？

纸巾品牌		
x	y	z
91	99	83
100	96	88
88	94	89
89	99	76

36. 美国环境保护署对全美各地城市空气中的污染物水平进行监测。臭氧污染水平使用一个取值范围为 0~500 的指数度量；指数分值越低表明健康风险越小，指数分值越高表明健康风险越大。四个城市（亚拉巴马州伯明翰、田纳西州孟菲斯市、阿肯色州小石城和密西西比州杰克逊）2012 年 10 天的臭氧污染峰值水平的数据如下表所示（U. S. EPA website，2012 年 3 月 20 日）。

日期	亚拉巴马州伯明翰	田纳西州孟菲斯市	阿肯色州小石城	密西西比州杰克逊
1 月 9 日	18	20	18	14
1 月 17 日	23	31	22	30
1 月 18 日	19	25	22	21
1 月 31 日	29	36	28	35
2 月 1 日	27	31	28	24
2 月 6 日	26	31	31	25
2 月 14 日	31	24	19	25
2 月 17 日	31	31	28	28
2 月 20 日	33	35	35	34
2 月 29 日	20	42	42	21

在 $\alpha = 0.05$ 的显著性水平下，检验四个城市之间的平均臭氧污染水平是否存在显著差异。

38. 对一个新产品提出三种不同的装配方法，选择使用完全随机化实验设计方法来确定哪一种装配方法能使每小时装配的产品数量最多。随机抽取了 30 名工人，并指定每个人使用三种装配方法中的一种。每名工人装配的产品数量如下表所示。

方法		
A	B	C
97	93	99
73	100	94
93	93	87
100	55	66
73	77	59

（续）

方法		
A	B	C
91	91	75
100	85	84
86	73	72
92	90	88
95	83	86

在 $\alpha=0.05$ 的显著性水平下，根据这些数据资料来检验：是否能够看出每种方法装配的产品平均件数是相同的。

40. 一个研究机构用油耗来检测三种品牌汽油的性能特性。因为对于不同品牌的汽车，汽油表现出的性能特征不同，因此我们选择了五种品牌的汽车，并且在实验中把五种品牌的汽车看成是区组；即每种品牌的汽车都用三种品牌的汽油进行测试。实验结果（单位：英里/加仑）如下表所示。

		汽油品牌		
		Ⅰ	Ⅱ	Ⅲ
汽车	A	18	21	20
	B	24	26	27
	C	30	29	34
	D	22	25	24
	E	20	23	24

a. 在 $\alpha=0.05$ 的显著性水平下，三种品牌汽油每加仑行驶的平均里程数存在着差异显著吗？

b. 利用完全随机化设计的 ANOVA 方法分析实验数据。将得到的结果与（a）中得到的结果进行比较。设法删除区组影响的优点是什么？

42. 美国职业棒球大联盟的特许经营权依赖于观看比赛的观众人数，因为门票收入占其总收入的相当大份额，并且周末比赛是特别重要的。2011 赛季的前三个月（4 月、5 月和 6 月），休斯敦太空人队与 7 支球队周末系列比赛的观众人数如下表所示（ESPN website，2012 年 1 月 12 日）。

比赛对手	星期五	星期六	星期日
佛罗里达马林鱼队	41 042	25 421	22 299
圣迭戈教士队	23 755	28 100	22 899
密尔沃基酿酒人队	25 734	26 514	23 908
纽约大都会队	28 791	31 140	28 406
亚利桑那响尾蛇队	21 834	31 405	21 882
亚特兰大勇士队	29 252	32 117	23 765
坦帕湾光芒队	26 682	27 208	23 965

在 $\alpha=0.05$ 的显著性水平下，检验这三天比赛的平均观众人数是否是相同的。休斯敦太空人队正在考虑，在赛季后半程的每场周末系列比赛运作一个特别的促销活动，以增加观众人数。上表数据是否表明有某一天，在这天休斯敦太空人队应该安排这些促销活动？

44. 一家制造公司设计了一个析因实验，以确定由两台机器生产的有缺陷的零部件数量是否有差异，以及确定生产的有缺陷的零部件数量是否还取决于每台机器的原材料的投料方式，是用人工投料，还是用自动系统投料。生产的有缺陷的零部件数量的统计资料如下表所示。在 $\alpha=0.05$ 的显著性水平下，检验机器、投料系统以及它们之间交互作用的显著影响。

	投料系统	
	人工	自动
机器 1	30	30
	34	26
机器 2	20	24
	22	28

案例 13-1　Wentworth 医疗中心

作为对 65 岁和 65 岁以上的老人长期研究的一部分，位于纽约州北部地区的 Wentworth 医疗中心的社会学家和内科医生调查了地理位置和抑郁症之间的关系。抽取出 60 名健康状况不错的人组成一个随机样本，其中 20 人居住在佛罗里达州，20 人居住在纽约州，20 人居住在北卡罗来纳州。对随机选中的每个人进行了一次测量抑郁症的标准化检验，收集到的数据如下表所示；较高的检验分数表示有较高程度的抑郁症。这些数据被存放在本书所附光盘名为 Medical1 的文件中。

研究的第二部分是考察地理位置与患有慢性病的 65 岁和 65 岁以上的老人得抑郁症之间的关系，这些慢性病诸如关节炎、高血压和心脏病等。具有这种身体状况的老人也被抽取出 60 人组成一个随机样本，同样 20 人居住在佛罗里达州，20 人居住在纽约州，20 人居住在北卡罗来纳州。这项研究所记录的抑郁症程度的数据如下表所示。这些数据被存放在本书所附光盘名为 Medical2 的文件中。

Medical1 的数据			Medical2 的数据		
佛罗里达州	纽约州	北卡罗来纳州	佛罗里达州	纽约州	北卡罗来纳州
3	8	10	13	14	10
7	11	7	12	9	12
7	9	3	17	15	15
3	7	5	17	12	18
8	8	11	20	16	12
8	7	8	21	24	14
8	8	4	16	18	17
5	4	3	14	14	8
5	13	7	13	15	14
2	10	8	17	17	16
6	6	8	12	20	18
2	8	7	9	11	17
6	12	3	12	23	19
6	8	9	15	19	15
9	6	8	16	17	13
7	8	12	15	14	14
5	5	6	13	9	11
4	7	3	10	14	12
7	7	8	11	13	13
3	8	11	17	11	11

管理报告

1. 利用描述统计学方法汇总这两部分研究的数据。关于抑郁症的得分，你的初步观测结果是什么？
2. 对于两个数据集使用方差分析方法，在每种情况下陈述需要进行检验的假设，你的结论是什么？
3. 在适当的地方使用单个处理方法的统计推断。你的结论是什么？

案例 13-2　销售人员的报酬

假设大旧金山地区一个销售专业人员的地方分会对该组织的成员进行了一项调查，其目的是：对于受雇的销售人员，无论销售场所是在室内还是在户外，如果他们的资历与年薪之间存在某种关系的话，则对这种关系进行研究。在调查中，受访者被要求在三种资历水平中指定其中的一种：低（1～10 年），中（11～20 年）和高（21 或 21 年以上）。调查所得数据的一部分如下所示。完整的数据集由 120 组观测值组成，它们被存放在本书所附光盘名为 SalesSalary 的文件中。

观测值	年薪（美元）	销售场所	工作资历	观测值	年薪（美元）	销售场所	工作资历
1	53 938	室内	中	10	54 768	室内	中
2	52 694	室内	中	⋮	⋮	⋮	⋮
3	70 515	户外	低	115	58 080	室内	高
4	52 031	室内	中	116	78 702	户外	中
5	62 283	户外	低	117	83 131	户外	中
6	57 718	室内	低	118	57 788	室内	高
7	79 081	户外	高	119	53 070	室内	中
8	48 621	室内	低	120	60 259	户外	低
9	72 835	户外	高				

管理报告

1. 用描述统计学汇总这些数据。

2. 不考虑销售人员的工作资历和工作场所，建立一个 95% 的全体销售人员平均年薪的置信区间估计。

3. 建立一个 95% 的室内销售人员平均年薪的置信区间估计。

4. 建立一个 95% 的户外销售人员平均年薪的置信区间估计。

5. 在 $\alpha=0.05$ 的显著性水平下，并且暂时不考虑销售人员工作资历的影响，利用方差分析方法，检验销售人员工作场所的显著差异。

6. 在 $\alpha=0.05$ 的显著性水平下，并且暂时不考虑销售人员工作场所的影响，利用方差分析方法，检验销售人员工作资历的显著差异。

7. 在 $\alpha=0.05$ 的显著性水平下，检验销售人员的工作场所、工作资历和交互作用的显著差异。

第 14 章

简单线性回归

CHAPTER 14

实践中的统计

联盟数据系统（ALLIANCE DATA SYSTEMS）[⊖]

得克萨斯州，达拉斯

在飞速发展的客户关系管理行业中，联盟数据系统（ADS）可为顾客提供交易代理、信贷服务和营销服务等一系列服务。ADS的顾客群集中在零售业、加油站/便利店、公用事业和交通运输业四大行业。1983年，Alliance开始为旗下的零售业务、加油站和餐饮业提供不间断的信贷业务服务；今天，该公司大约有6 500多名员工为世界各地的顾客提供这些服务。ADS仅在美国就有140 000台零售点终端机，每年处理超过25亿宗交易业务。ADS代理了49种贴有零售商标签的服务计划，差不多有7 200万消费者持有参加这些计划的优待卡，这使得该公司在全美贴有零售商标签的信用服务行业中排名次席。基于良好的业绩，2001年，ADS首次公开发行股票并在纽约证券交易所成功上市。

作为营销服务的手段之一，ADS设计了直接向顾客邮寄宣传品的促销活动。由于它的数据库储存了1亿多名顾客消费习惯的信息，所以ADS把那些最有可能的顾客作为促销目标，通过直接向他们邮寄宣传品达到获得收益的目的。公司的分析发展部门运用回归分析方法，建立能度量并预测顾客对促销活动反应的模型。一些回归模型预测了顾客收到促销宣传品后购买商品的概率，另一些回归模型则预测了这些顾客购买商品所花费的金额。

在某一项特定的促销活动中，零售连锁店的目标是吸引新顾客。为了预测此项促销活动的效果，ADS的分析师们从顾客信息数据库中选取了一个样本，向样本中的每一位顾客发放了促销宣传材料，然后将样本顾客对此项活动反馈的数据收集起来并加以处理。样本数据不但包括顾客由于促销活动而购买商品的金额，而且还包括了各种各样的、有助于预测销售额的顾客特定变量。顾客特定变量是指顾客在过去39个月里从相关商店中赊购商品的总金额，该变量对预测顾客购买商品的金额将会起到很大的作用。ADS的分析师们建立了购买商品的总金额与过去从相关商店中赊购商品的金额之间关系的估计的回归方程：

$$\hat{y} = 26.7 + 0.002\,05x$$

式中，$\hat{y}$代表购买商品的总金额；x代表过去从相关商店中赊购商品的金额。

利用这一方程，我们能够预测出：一位在过去39个月里从相关商店中赊购了10 000美元的顾客，对于直接向他们邮寄宣传品的反应将是消费47.20美元。在本章中，你将学会如何建立这种类型的估计的回归方程。

为了提高前述方程的预测能力，ADS的分析师建立的最终模型还包含了一些其他变量。比如，是否拥有一家银行的信用卡、估计的收入水平以及每次光顾被挑选出来的商店的平均消费额度。在下一章里，我们将学习如何把这些额外的变量整合到多元回归模型中。

管理决策，经常取决于对两个或多个变量之间关系的分析。例如，一位市场销售经理权衡了广告费用和销售收入之间的关系后，才可能尝试去预测一个给定水平的广告费用，能带来多少销售收入。又如，一家公用事业公司可以利用白天最高气温与电力需求之间的关系，根据下个月白天最高气温的预报，来预测下个月的用电量。通常，一位管理人员要依靠直觉来判断两个变量的关系。但是，如果能取得数据，我们就能利用统计方法来建立一个表示变量之间相互关系的方程，我们将这一统计方法称为回归分析。

在回归术语中，我们把被预测的变量称为**应变量**（dependent variable），把用来预测应变量值的一个或多个变量称为**自变量**（independent variable）。例如，在分析广告费用对销售收入的影响时，市场销售经理想要预测销售收入的愿望使我们想到，应该用销售收入作为应变量，而用来帮助预测销售收入的广告费用应作为自变量。在统

⊖ 作者感谢联盟数据系统分析发展部主任菲利普·克莱曼斯先生，他为“实践中的统计”提供了本案例。

计符号上，y 代表应变量，x 代表自变量。

在本章中，我们仅仅讨论最简单类型的回归分析，它只包括一个自变量和一个应变量，二者之间的关系可以用一条直线近似表示。这种回归分析被称为**简单线性回归**（simple linear regression）。包括两个或两个以上自变量的回归分析被称为多元回归分析；多元回归和涉及曲线关系的情形，将在第 15 章和第 16 章中介绍[⊖]。

14.1 简单线性回归模型

Armand 比萨饼连锁店是经营意大利食品的餐馆，它们分布在美国 5 个州的范围内。Armand 比萨饼连锁店的最佳位置是在大学校园附近。管理人员确信，这些连锁店的季度销售收入（用 y 表示）与学生人数（用 x 表示）是正相关的。也就是说，学生较多的校园附近的连锁店比学生较少的校园附近的连锁店，有获得较大的季度销售收入的倾向。利用回归分析，我们能求出一个说明应变量 y 是如何依赖自变量 x 的方程。

14.1.1 回归模型和回归方程

在 Armand 比萨饼连锁店的例子中，总体是由所有的 Armand 比萨饼连锁店组成的。对于总体中的每一个连锁店，都有一个 x 值（学生人数）和一个对应的 y 值（季度销售收入）。描述 y 如何依赖于 x 和误差项的方程称为**回归模型**（regression model）。下面是用于简单线性回归的回归模型。

简单线性回归模型

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{14-1}$$

式中，β_0 和 β_1 称为模型的参数；ε 是一个随机变量，称为模型的误差项。误差项说明了包含在 y 里面但不能被 x 和 y 之间的线性关系解释的变异性。

Armand 比萨饼连锁店总体还可以视为由若干个子总体组成的集合，每一个子总体都对应一个不同的 x 的值。例如，一个子总体是由有 8 000 名学生的校园附近的所有 Armand 比萨饼连锁店组成的；另一个子总体是由有9 000 名学生的校园附近的所有 Armand 比萨饼连锁店组成的，等等。每一个子总体都对应一个 y 值的分布。于是，位于有 8 000 名学生的校园附近的连锁店对应一个 y 值的分布，位于有 9 000 名学生的校园附近的连锁店对应另一个 y 值的分布，等等。y 值的每一个分布都有它自己的平均值或期望值。描述 y 的期望值 $E(y)$ 如何依赖于 x 的方程称为**回归方程**（regression equation）。对于简单线性回归情形，回归方程如下。

简单线性回归方程

$$E(y) = \beta_0 + \beta_1 x \tag{14-2}$$

简单线性回归方程的图形是一条直线；β_0 是回归直线的 y 轴截距，β_1 是斜率，对于一个给定的 x 值，$E(y)$ 是 y 的平均值或期望值。

对于简单线性回归方程，它所代表的各种可能的回归线的实例如图 14-1 所示。图 14-1a 中的回归线表示 y 的平均值与 x 正相关，较大的 x 值，对应的 $E(y)$ 的值也较大。图 14-1b 中的回归线表示 y 的平均值与 x 负相关，较大的 x 值，对应的 $E(y)$ 的值则较小。图 14-1c 中的回归线表示 y 的平均值与 x 无关，即对于 x 的每一个值，y 的平均值 $E(y)$ 都是相同的。

⊖ 最先应用统计方法来研究两个变量之间关系的是弗朗西斯・高尔顿（Francis Galton，1822—1911）。他对父子身高之间的关系很感兴趣，并致力于此方面的研究。高尔顿的追随者，卡尔・皮尔逊（Karl Pearson，1857—1936）则通过 1 078 对受试者，分析了父子身高之间的关系。

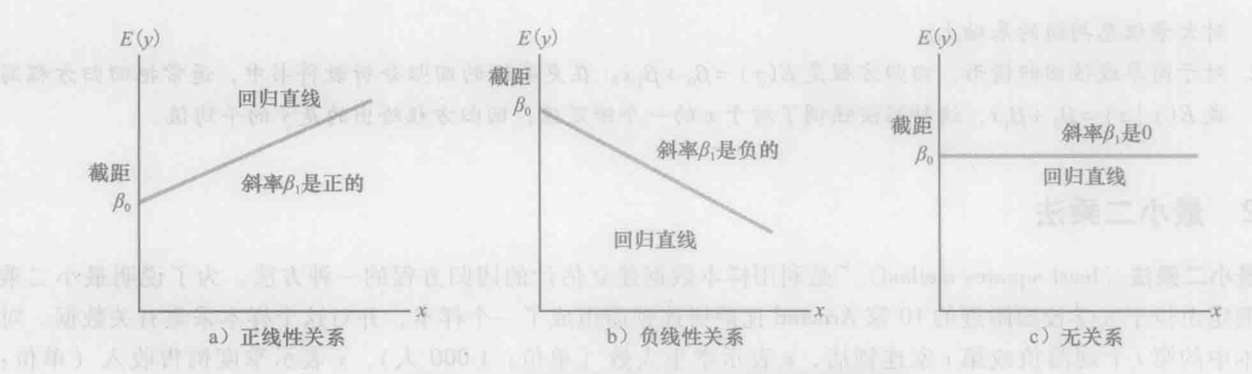


图 14-1 在简单线性回归中的各种可能的回归线

14.1.2 估计的回归方程

如果总体参数 β_0 和 β_1 的值是已知的，那么对于一个给定的 x 值，我们能利用式（14-2）计算 y 的平均值。遗憾的是，在实际中参数 β_0 和 β_1 的值常常是未知的，我们必须利用样本数据去估计它们。我们计算样本统计量 b_0 和 b_1 作为总体参数 β_0 和 β_1 的估计量。用样本统计量 b_0 和 b_1 替代回归方程中的未知参数 β_0 和 β_1，我们得到了**估计的回归方程**（estimated regression equation）。对于简单线性回归情形，估计的回归方程如下。

估计的简单线性回归方程

$$\hat{y} = b_0 + b_1 x \tag{14-3}$$

图 14-2 给出了简单线性回归估计步骤的一个概要。[1]

估计的简单线性回归方程的图形，被称作估计的回归线；b_0 是 y 轴截距，b_1 是斜率。在下一节，我们将说明如何利用最小二乘法计算估计的回归方程中的 b_0 和 b_1 的值。

通常，对于 x 的一个给定值，$\hat{y}$是 y 的平均值 $E(y)$ 的一个点估计。于是，为了估计位于有 10 000 名学生的校园附近的所有 Armand 比萨饼连锁店季度销售收入的平均值或期望值，我们只需要将 10 000 代入式（14-3）中的 x 即可。但是，在某些情形下，Armand 比萨饼连锁店的管理人员可能对预测某一家特定的连锁店的销售收入更感兴趣。例如，假定Armand比萨饼连锁店的管理人员希望预测位于有 10 000 名学生的 Talbot 大学附近的一家连锁店的季度销售收入。后面我们将会证明，对于 x 的一个给定值，y 的最优估计仍然由$\hat{y}$给出。于是，为了预测位于 Talbot 大学附近的这家比萨饼连锁店的季度销售收入，Armand 比萨饼连锁店的管理人员仍然需要将10 000代入式（14-3）中的 x。[2]

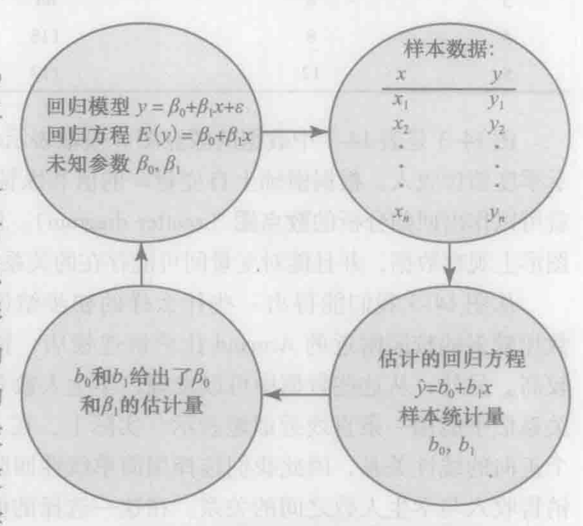


图 14-2 简单线性回归的估计步骤

注释和评论

1. 我们不能把回归分析看作在变量之间建立一个因果关系的过程。回归分析只能表明，变量是如何或者是以怎样的程度彼此联系在一起的。变量之间有关因果关系和相互影响的任何结论，必须建立在人们在应用时

[1] β_0 和 β_1 的估计是一个统计过程，这个过程与在第 7 章中讨论过的 μ 的估计非常相似。β_0 和 β_1 是我们感兴趣的未知参数，b_0 和 b_1 是用于估计未知参数的样本统计量。

[2] 对于 x 的一个给定值，$\hat{y}$给出了两个值：一个是 y 的平均值 $E(y)$ 的一个点估计；一个是 y 的一个个别值的预测值。

对大量信息判断的基础上。

2. 对于简单线性回归情形，回归方程是 $E(y)=\beta_0+\beta_1 x$。在更高级的回归分析教科书中，通常把回归方程写成 $E(y|x)=\beta_0+\beta_1 x$，这种写法强调了对于 x 的一个给定值，回归方程给出的是 y 的平均值。

14.2 最小二乘法

最小二乘法（least squares method）[㊀]是利用样本数据建立估计的回归方程的一种方法。为了说明最小二乘法，假定由位于大学校园附近的 10 家 Armand 比萨饼连锁店组成了一个样本，并对这个样本采集有关数据。对于样本中的第 i 个观测值或第 i 家连锁店，x_i 表示学生人数（单位：1 000 人），y_i 表示季度销售收入（单位：1 000美元）。样本中 10 家 Armand 比萨饼连锁店的 x_i 和 y_i 的数值如表 14-1 所示。我们看到，对于第 1 家连锁店，$x_1=2$，$y_1=58$，表示这家连锁店位于有 2 000 名学生的校园附近，它的季度销售收入为 58 000 美元。对于第 2 家连锁店，$x_2=6$，$y_2=105$，表示这家连锁店位于有 6 000 名学生的校园附近，它的季度销售收入为 105 000美元。销售收入最多的连锁店是第 10 家连锁店，它位于有 26 000 名学生的校园附近，季度销售收入为 202 000 美元。

表 14-1 10 家 Armand 比萨饼连锁店的学生人数和季度销售收入数据

连锁店 i	学生人数（1 000 人）x_i	销售收入（1 000 美元）y_i	连锁店 i	学生人数（1 000 人）x_i	销售收入（1 000 美元）y_i
1	2	58	6	16	137
2	6	105	7	20	157
3	8	88	8	20	169
4	8	118	9	22	149
5	12	117	10	26	202

图 14-3 是表 14-1 中数据的散点图。横轴表示学生人数，纵轴表示季度销售收入。根据横轴上自变量 x 的值和纵轴上应变量 y 的值，就可以作出回归分析的**散点图**（scatter diagram）。散点图使我们能从图形上观察数据，并且能对变量间可能存在的关系得出初步的结论。

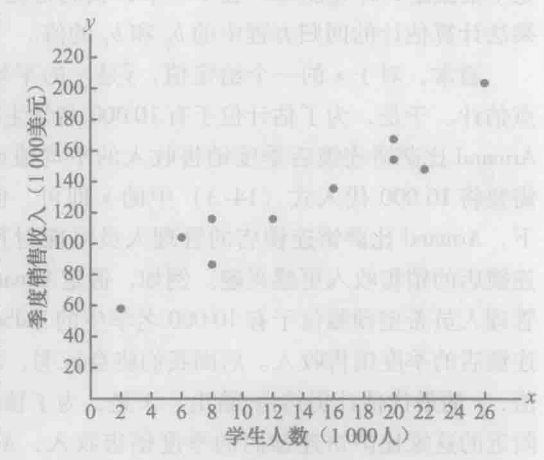


图 14-3 Armand 比萨饼连锁店的学生人数和季度销售收入的散点图

从图 14-3 我们能得出一些什么样的初步结论呢？位于学生人数比较多的校园附近的 Armand 比萨饼连锁店，销售收入似乎也比较高。另外，从这些数据中可以发现，学生人数和销售收入之间的关系似乎能用一条直线近似地表示。实际上，在 x 和 y 之间存在一个正向的线性关系，因此我们选择用简单线性回归模型来表示季度销售收入与学生人数之间的关系。在这一选择的假定下，我们接下来的任务是利用表 14-1 中的样本数据，去确定估计的简单线性回归方程中的 b_0 和 b_1 的值。对于第 i 家 Armand 比萨饼连锁店，估计的简单线性回归方程是

$$\hat{y}_i = b_0 + b_1 x_i \tag{14-4}$$

式中，$\hat{y}_i$ 代表第 i 家连锁店季度销售收入的估计值（1 000 美元）；b_0 代表估计的回归直线的 y 轴截距；b_1 代表估计的回归直线的斜率；x_i 代表第 i 家连锁店的学生人数（1 000 人）。

用 y_i 表示第 i 家连锁店季度销售收入的观测（实际）值，式（14-4）中 $\hat{y}_i$ 表示第 i 家连锁店季度销售收入的预测值，样本中的每一家连锁店都将有的一个季度销售收入的观测值 y_i 和一个季度销售收入的预测值 $\hat{y}_i$。为了使

㊀ 在简单线性回归中，每一对观测值由两个值组成，一个是自变量的观测值，一个是应变量的观测值。

估计的回归直线能对样本数据有一个好的拟合，我们希望季度销售收入的观测值与季度销售收入的预测值之间的差要小。

最小二乘法[1]是利用样本数据，通过使应变量的观测值 y_i 与应变量的预测值 $\hat{y}_i$ 之间的离差平方和达到最小的方法求得 b_0 和 b_1 的值。最小二乘法准则由式（14-5）给出。

最小二乘法准则

$$\min \sum (y_i - \hat{y}_i)^2 \tag{14-5}$$

式中，y_i 为对于第 i 次观测，应变量的观测值；$\hat{y}_i$ 为对于第 i 次观测，应变量的预测值。

微分学可以证明（见附录14A），使式（14-5）达到最小的 b_0 和 b_1 的值能利用式（14-6）和式（14-7）求得。

估计的回归方程的斜率和 y 轴截距

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \text{[2]} \tag{14-6}$$

$$b_0 = \bar{y} - b_1\bar{x} \tag{14-7}$$

式中，x_i 为对于第 i 次观测，自变量的观测值；y_i 为对于第 i 次观测，应变量的观测值；$\bar{x}$ 为自变量的样本平均值；$\bar{y}$ 为应变量的样本平均值；n 为总观测次数。

在 Armand 比萨饼连锁店的例子中，应用最小二乘法估计回归方程的一些必要的计算在表 14-2 中列出。由于样本是由 10 家 Armand 比萨饼连锁店组成，所以我们有观测次数 $n = 10$。因为式（14-6）和式（14-7）需要 $\bar{x}$ 和 $\bar{y}$，我们的工作从计算 $\bar{x}$ 和 $\bar{y}$ 开始。

表 14-2 Armand 比萨饼连锁店的最小二乘估计的回归方程的计算

连锁店 i	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
合计	140	1 300			2 840	568
	$\sum x_i$	$\sum y_i$			$\sum(x_i-\bar{x})(y_i-\bar{y})$	$\sum(x_i-\bar{x})^2$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{1\,300}{10} = 130$$

[1] 德国数学家高斯提出了最小二乘法。

[2] 计算 b_1 的另一个公式是

$$b_1 = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n}$$

当利用计算器计算 b_1 时，通常推荐用公式（14-6）。

当利用计算器计算 b_1 时，中间计算过程应尽可能地多地保留有效数字。我们建议，至少应保留四位有效数字。

利用式（14-6）和式（14-7），以及表14-2中的资料，我们能计算Armand比萨饼连锁店的估计的回归方程的斜率和y轴截距。斜率b_1的计算过程如下

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{2\,840}{568} = 5$$

y轴截距b_0的计算如下

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5 \times 14 = 60$$

于是，估计的回归方程是

$$\hat{y} = 60 + 5x$$

这一方程在散点图上的图形，如图14-4所示。估计的回归方程的斜率（$b_1 = 5$）是正的，也就是说，随着学生人数的增加，Armand比萨饼连锁店的季度销售收入也增加。实际上，我们可以得出结论，学生人数每增加1 000人，Armand比萨饼连锁店期望增加的季度销售收入将是5 000美元；即每增加1名学生，期望增加销售收入5美元（因为销售收入的计量单位是1 000美元，学生人数的计量单位是1 000人）。

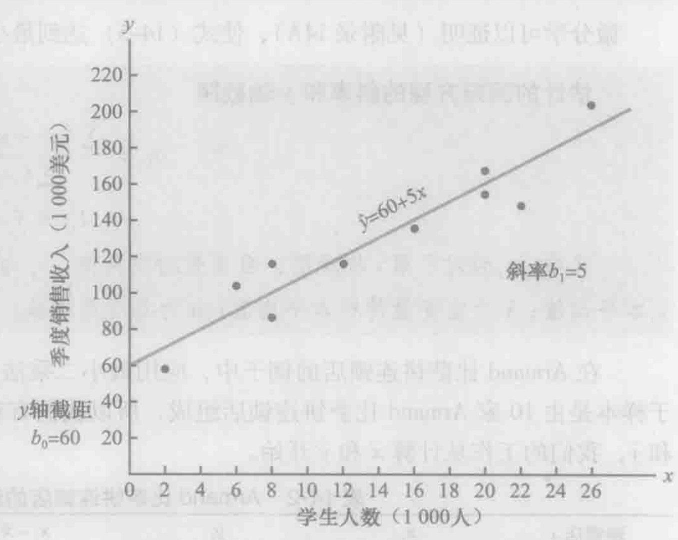


图14-4 Armand比萨饼连锁店估计的回归方程：$\hat{y} = 60 + 5x$的图示

如果我们相信，应用最小二乘法得到的估计的回归方程能充分地描述两个变量x和y之间的关系，那么对于一个已知的x值，利用估计的回归方程去预测y的值将被认为是合理的。例如，有一家连锁店，它位于有16 000名学生的校园附近，如果我们想要预测这家连锁店的季度销售收入，那么我们应当计算

$$\hat{y} = 60 + 5 \times 16 = 140$$

所以，对于这家Armand比萨饼连锁店，我们预测的季度销售收入是140 000美元。在下一节，我们将讨论利用估计的回归方程进行估计和预测合理性的评价方法。⊖

注释和评论

最小二乘法通过使应变量的观测值y_i与应变量的预测值$\hat{y}_i$之间的离差平方和达到最小，得到了估计的回归方程。最小二乘法准则就是选择能与样本数据有最佳拟合方程的准则。如果利用某些其他准则，例如最小化y_i和$\hat{y}_i$之间的绝对离差的和，我们将得到一个不同的方程。在实践中，最小二乘法是应用最广泛的方法。

练习

方法

2. 已知两变量x和y的5组观测值。

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a. 绘制出这些数据的散点图。

b. 根据在（a）中作出的散点图，这两个变量之间显示出什么关系？

c. 尝试着画一条穿过这些数据的直线，来近似x和y之间的关系。

d. 利用式（14-6）和式（14-7），计算b_0和b_1的值，建立估计的回归方程。

e. 当$x = 10$时，利用估计的回归方程，预测y的值。

⊖ 在自变量取值范围以外，利用估计的回归方程进行预测时要特别谨慎。因为在这个范围以外，我们不能保证变量之间存在同样的关系。

应用

☆4. 下表给出的是在五家零售贸易业公司中工作的女性职工所占的百分比，以及在每一家公司中由女性担任管理工作的百分比数据。

女性职工的百分比（%）	67	45	73	54	61
女性担任管理工作的百分比（%）	49	21	65	47	33

a. 以公司中女性职工所占的百分比为自变量，绘制出这些数据的散点图。

b. 根据在（a）中作出的散点图，这两个变量之间显示出什么关系？

c. 尝试在公司中女性职工所占的百分比和公司中女性担任管理工作的百分比之间建立一个近似的关系。

d. 计算 b_0 和 b_1 的值，建立估计的回归方程。

e. 如果一家公司的女性职工占60%，试预测该公司由女性担任管理工作的百分比。

6. 美国国家橄榄球联盟记录了每名球员和每支球队各种表现的数据。为了调查传球对于一支球队能否获胜的重要作用，由10支美国国家橄榄球联盟的球队组成一个随机样本，每支球队在2011年赛季每次进攻的平均传球码数（Yds/Att）和比赛的获胜率（WinPct）数据如下表所示（NFL website，2012年2月12日）

球队名称	Yds/Att	WinPct
Arizona Cardinals	6.5	50
Atlanta Falcons	7.1	63
Carolina Panthers	7.4	38
Chicago Bears	6.4	50
Dallas Cowboys	7.4	50
New England Patriots	8.3	81
Philadelphia Eagles	7.4	50
Seattle Seahawks	6.1	44
St. Louis Rams	5.2	13
Tmpa Bay Buccaneers	6.2	25

a. 用水平轴表示平均每次传球码数，用纵轴表示比赛的获胜率，绘制出这些数据的散点图。

b. 根据在（a）中作出的散点图，这两个变量之间显示出什么关系？

c. 建立估计的回归方程，使这个方程在球队每次进攻的平均传球码数已知时，能用来预测球队的获胜率。

d. 对估计的回归方程的斜率作出解释。

e. 如果2011年赛季，Kansas City Chiefs队的平均每次传球码数是6.2，利用在（c）中建立的估计的回归方程预测该队的获胜率。（注：在2011年赛季，Kansas City Chiefs队的比赛记录是7胜9负。）将你的预测结果与Kansas City Chiefs队的实际获胜率进行比较。

8. 美国个人投资者协会的在线折扣经纪商调查，是根据协会会员与折扣经纪商的经验进行的一项调查工作。作为调查的一部分，要求AAII的会员对经纪商的执行速度能力作出评估，以及对电子交易提供一个整体满意度等级。可能的回答（分数）分别为无意见（0），不满意（1），比较满意（2），满意（3），非常满意（4）。根据每个受访者提供的分数，计算出加权平均数，得到每位经纪商的总分数。部分调查结果如下表所示（AAII website，2012年2月7日）。

经纪公司	速度	满意度
Scottrade，Inc.	3.4	3.5
Charles Schwab	3.3	3.4
Fidelity Brokerage Services	3.4	3.9
TD Ameritrade	3.6	3.7
E * Trade Financial	3.2	2.9
Vanguard Brokerage Services	3.8	2.8
USAA Brokerage Services	3.8	3.6
Thinkorswim	2.6	2.6
Wells Fargo Investments	2.7	2.3
Interactive Brokers	4.0	4.0
Zecco.com	2.5	2.5

a. 以执行速度为自变量，绘制出这些数据的散点图。

b. 根据在（a）中作出的散点图，这两个变量之间显示出什么关系？

c. 利用最小二乘法，建立估计的回归方程。

d. 对估计的回归方程的斜率作出解释。

e. 假定Zecco.com开发了新软件提高了它们执行速度的等级。如果新软件能将执行速度的等级从目前的2.5提高到其他10个被调查的经纪公司的平均执行速度等级，请你预测Zecco.com的整体满意度等级是多少？

10. 2009年3月31日，福特汽车公司的股票成交价是2.63美元，为26年来新低。福特汽车公司的董事会将估价为1 600万美元的优先购股权和受限制股票给予CEO作为报酬的补偿。2011年4月26日，

福特汽车公司的股票价格增长到15.58美元，CEO优先购股权的价值已经增加到2.028亿美元，增值收益为1.868亿美元。下表是10家公司2009年和2011年的股票价格，2008年、2009年年底股票期权和授予CEO的购股权，以及2011年股票期权和购股权的价值。下表还给出了股票价格增长的百分比和购股权价值的百分比涨幅（《华尔街日报》，2011年4月27日）

a. 以股票价格增长的百分比为自变量，绘制出这些数据的散点图。

b. 根据在（a）中作出的散点图，这两个变量之间显示出什么关系？

c. 利用最小二乘法，建立估计的回归方程。

d. 对估计的回归方程的斜率作出解释。

e. 以股票价格为衡量标准，给予CEO的报酬是否体现了是基于业绩的增长？

公司	2009年股价（美元）	2011年股价（美元）	股价增长（%）	2009年股票期权和购股权的价值（100万美元）	2011年股票期权和购股权的价值（100万美元）	购股权价值的涨幅（%）
Ford Motor	2.63	15.58	492	16.0	202.8	1 168
Abercrombie & Fitch	23.80	70.47	196	46.2	196.1	324
Nabors Industries	9.99	32.06	221	37.2	132.2	255
Starbucks	9.99	32.06	221	12.4	75.9	512
Salesforce. com	32.73	137.61	320	7.8	67.0	759
Starwood Hotels	12.70	60.28	375	5.8	57.1	884
Caterpillar	27.96	111.94	300	4.0	47.5	1 088
Oracle	18.70	34.97	94	61.9	97.5	58
Capital One	12.24	54.61	346	6.0	40.6	577
Dow Chemical	8.43	39.97	374	5.0	38.8	676

12. Concur Technologies公司是一家位于华盛顿州Redmond的大型费用管理公司。《华尔街日报》要求公司对830万份的费用报告进行调查，以提供有关商务旅行费用的意见。公司的数据分析表明纽约是最昂贵的城市，平均每天酒店房租为198美元，平均娱乐费用为172美元，其中包括团队的餐饮、演出和体育比赛的门票以及其他活动。相比之下，这两项支出在美国的平均费用是，酒店房租为89美元，娱乐费用为99美元。从25个参访人数最多的美国城市中抽取9个城市组成一个随机样本，这9个城市的平均每天酒店房租和平均娱乐费用如下表所示（《华尔街日报》，2011年8月18日）。

城市	酒店房租（美元）	娱乐费用（美元）
波士顿	148	161
丹佛	96	105
纳什维尔	91	101
新奥尔良	110	142
菲尼克斯	90	100
圣迭戈	102	120
旧金山	136	167
圣何塞	90	140
坦帕	82	98

a. 以酒店房租为自变量，绘制出这些数据的散点图。

b. 根据在（a）中作出的散点图，酒店房租和娱乐费用这两个变量之间显示出什么关系？

c. 利用最小二乘法，建立估计的回归方程。

d. 对估计的回归方程的斜率作出解释。

e. 芝加哥平均每天酒店房租是128美元，较美国的平均水平高。预测芝加哥每天的平均娱乐费用。

14. *PC World*杂志根据外观、性能、设计和价格等4个方面的特性对10台超薄型便携式笔记本电脑进行等级评估。每个特性的等级评估采用用百分制。对于每一台超薄型便携式笔记本电脑，将得到一个称为*PCW World*的总体等级评估分。下表是10台超薄型便携式笔记本电脑的外观和*PCW World*的等级评估分（*PC World* website，2009年2月5日）

a. 以外观等级评估分为自变量，绘制出这些数据的散点图。

b. 根据在（a）中作出的散点图，外观等级评估分和*PCW World*等级评估分这两个变量之间显示出什么关系？

c. 利用最小二乘法，建立估计的回归方程。

d. 对于一台新的超薄型便携式笔记本电脑，它的外观等级评估分为70，预测该台电脑的*PCW World*等级评估分。

笔记本电脑型号	外观等级评估分	*PCW World* 等级评估分	笔记本电脑型号	外观等级评估分	*PCW World* 等级评估分
Thinpad X200	87	83	Thinpad X300	76	78
VGN-Z598U	85	82	Ideapad U110	81	77
U6V	80	81	Micro Express JFT2500	73	75
Elitebook 2530P	75	78	Toughbook W7	79	73
X360	80	78	HP Voodoo Envy133	68	72

14.3 判定系数

对 Armand 比萨饼连锁店的例子，我们求出了估计的回归方程是$\hat{y}=60+5x$，并用这个方程作为学生人数 x 和季度销售收入 y 之间线性关系的一个近似。现在摆在我们面前的问题是：估计的回归方程是否很好地拟合了样本数据？在这一节，我们将说明，**判定系数**（coefficient of determination）为估计的回归方程提供了一个拟合优度的度量。

对于样本中的第 i 次观测值，应变量的观测值 y_i 和应变量的预测值$\hat{y}_i$ 之间的离差称为**第 i 个残差**（i th residual）。第 i 个残差表示用$\hat{y}_i$ 去估计 y_i 的误差。于是，对于第 i 次观测值，它的残差是 $y_i-\hat{y}_i$。这些残差或误差的平方和是一个用最小二乘法来极小化的量。这个量也称为误差平方和，记作 SSE。

误差平方和

$$\text{SSE}=\sum(y_i-\hat{y}_i)^2 \tag{14-8}$$

在应用估计的回归方程去预测样本中应变量的值时将产生一个误差，SSE 的数值就是对这一误差的度量。

对于 Armand 比萨饼连锁店的例子，我们在表 14-3 中给出了计算误差平方和所需要的计算过程。例如，对于第 1 家连锁店，自变量的值是 $x_1=2$，应变量的值是 $y_1=58$。利用估计的回归方程，我们求出第 1 家连锁店季度销售收入的预测值是$\hat{y}_1=60+5\times2=70$。于是，对于第 1 家连锁店，用$\hat{y}_1$去预测 y_1 的残差是 $y_1-\hat{y}_1=58-70=-12$。误差的平方是 $(-12)^2=144$，我们将它写在表 14-3 的最后一列。对样本中的每一家连锁店计算残差并平方后，我们对它们求和得到 $SSE=1\,530$。于是，对于 Armand 比萨饼连锁店的例子，SSE = 1 530 度量了用估计的回归方程$\hat{y}=60+5x$ 预测季度销售收入所产生的误差。

表 14-3 Armand 比萨饼连锁店 SSE 的计算

连锁店 i	x_i = 学生人数（1 000 人）	y_i = 季度销售收入（1 000 美元）	季度销售收入预测值 $\hat{y}_i=60+5x_i$	误差 $y_i-\hat{y}_i$	误差平方 $(y_i-\hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SSE = 1 530

现在假定在不知道学生人数的情况下，要求我们给出季度销售收入的一个估计值。在没有任何相关变量信息的情况下，对任意一个给定的连锁店，我们可以利用样本平均值作为这家连锁店季度销售收入的一个估计值。表14-2表明，对于销售收入的样本数据，有$\sum y_i = 1\,300$。于是，对于由10家Armand比萨饼连锁店组成的样本，季度销售收入的平均值是$\bar{y} = \sum y_i / n = 1\,300/10 = 130$。在利用样本平均值$\bar{y} = 130$预测样本中每一家连锁店的季度销售收入将产生离差，这些离差的平方和我们列在表14-4的最后一列。对于样本中的第i家连锁店，离差$y_i - \bar{y}$给出了利用样本平均值$\bar{y}$去预测季度销售收入所产生的误差的一个度量。这些离差对应的平方和被称为总的平方和，记作SST。

总的平方和

$$SST = \sum (y_i - \bar{y})^2 \tag{14-9}$$

在表14-4中，最后一列底部的和就是Armand比萨饼连锁店例子的总平方和，它是SST = 15 730。

在图14-5中，我们给出了估计的回归线$\hat{y} = 60 + 5x$和对应于$\bar{y} = 130$的直线的图形。从图上我们注意到，样本点集聚在估计的回归线周围比集聚在直线$\bar{y} = 130$周围更紧密。例如，对于样本中的第10家连锁店，我们看到用$\bar{y} = 130$作为y_{10}的预测值比用$\hat{y}_{10} = 60 + 5 \times 26 = 190$作为$y_{10}$的预测值产生的误差大得多。我们可以把SST看作是观测值在直线$\bar{y} = 130$周围集聚程度的度量，而把SSE看作是观测值在回归线$\hat{y} = 60 + 5x$周围集聚程度的度量。[㊀]

为了度量在估计的回归线上的$\hat{y}$值与直线$\bar{y}$的偏离有多大，我们需要计算另一个平方和。这个平方和被称为回归平方和，记作SSR。

表14-4 Armand比萨饼连锁店SST的计算

连锁店 i	x_i = 学生人数（1 000人）	y_i = 季度销售收入（1 000美元）	离差 $y_i - \bar{y}$	离差平方 $(y_i - \bar{y})^2$
1	2	58	−72	5 184
2	6	105	−25	625
3	8	88	−42	1 764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1 521
9	22	149	19	361
10	26	202	72	5 184
				SST = 15 730

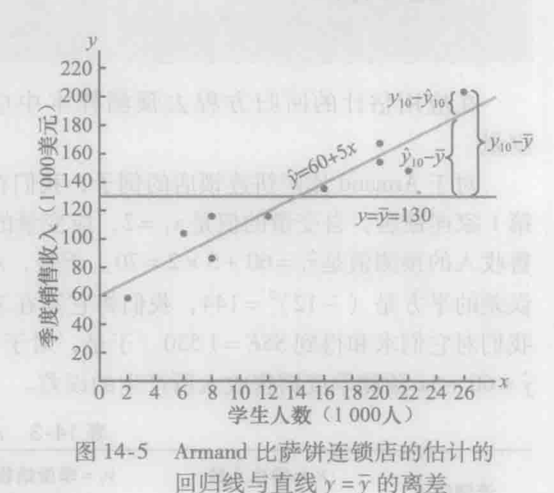


图14-5 Armand比萨饼连锁店的估计的回归线与直线$y = \bar{y}$的离差

回归平方和

$$SSR = \sum (\hat{y}_i - \bar{y})^2 \tag{14-10}$$

从上述讨论中，我们应该预期到SST，SSR和SSE三者之间是有联系的。事实上，这三个平方和之间的关系给出了统计学中最重要的一个结果。

SST，SSR和SSE之间的关系[㊁]

$$SST = SSR + SSE \tag{14-11}$$

式中，SST代表总的平方和；SSR代表回归平方和；SSE代表误差平方和。

㊀ 因为SST = 15 730，SSE = 1 530，所以，估计的回归线对样本数据的拟合比直线$y = \bar{y}$更好。

㊁ 我们把SSR理解为SST的被解释的部分，把SSE理解为SST的未被解释的部分。

式（14-11）表明，总的平方和能被分解成两部分：回归平方和和误差平方和。因此，如果已知这三个平方和中的任意两个，都能很容易地计算出第三个平方和。例如，在 Armand 比萨饼连锁店的例子中，我们已经知道 SSE = 1 530 和 SST = 15 730，所以从式（14-11）中解出 SSR，我们得到回归平方和是

$$SSR = SST - SSE = 15\,730 - 1\,530 = 14\,200$$

现在让我们看一下，如何利用这三个平方和 SST、SSR 和 SSE 为估计的回归方程给出一个拟合优度的度量。如果应变量的每一个值 y_i 都刚好落在估计的回归线上，那么估计的回归方程将给出一个完全的拟合。在这种情况下，对于每一个观测值，$y_i - \hat{y}_i$ 将等于零，从而导致 SSE = 0。因为 SST = SSR + SSE，我们看到对于一个完全拟合，SSR 必须等于 SST，并且比值 SSR/SST 必须等于 1。比较差的拟合将导致 SSE 的值比较大。从式（14-11）中解出 SSE，我们得到 SSE = SST − SSR。因此，当 SSR = 0，从而 SSE = SST 时，这时候 SSE 的值最大（即最差的拟合）。

比值 SSR/SST 将在 0 和 1 之间取值，我们利用这个比值对估计的回归方程的拟合优度作出评估。这个比值被称为判定系数，记作 r^2。

判定系数

$$r^2 = \frac{SSR}{SST} \tag{14-12}$$

对于 Armand 比萨饼连锁店的例子，判定系数的值是

$$r^2 = \frac{SSR}{SST} = \frac{14\,200}{15\,730} = 0.902\,7$$

如果我们用一个百分数表示判定系数，我们能把 r^2 理解为总平方和中能被估计的回归方程解释的百分比。对于 Armand 比萨饼连锁店的例子，在用估计的回归方程去预测季度销售收入时，我们能断定，总平方和中的 90.27% 能被估计的回归方程 $\hat{y} = 60 + 5x$ 所解释。换句话说，季度销售收入变异性的 90.27% 能被学生人数和销售收入之间的线性关系所解释。对于估计的回归方程，我们应该对得到一个这样好的拟合效果而感到满意。

相关系数

在第 3 章中，作为两变量 x 和 y 之间线性关系强度的描述性度量，我们介绍了**相关系数**（correlation coefficient）的概念。相关系数的数值总是介于 −1 ~ +1。若相关系数的数值等于 +1，则表示两个变量 x 和 y 之间存在完全正向的线性关系，即全部数据点都落在一条斜率为正的直线上；若相关系数的数值等于 −1，则表示两个变量 x 和 y 之间存在完全负向的线性关系，即全部数据点都落在一条斜率为负的直线上；若相关系数的数值接近于零，则表示两个变量 x 和 y 之间不存在线性关系。

在第 3.5 节中，我们曾经给出过计算样本相关系数的公式。如果我们已经完成了回归分析，并且计算出了判定系数 r^2，那么我们就能用下面的公式计算样本相关系数。

样本相关系数

$$r_{xy} = (b_1\text{ 的符号})\sqrt{\text{判定系数}} = (b_1\text{ 的符号})\sqrt{r^2} \tag{14-13}$$

式中，b_1 为估计的回归方程 $\hat{y} = b_0 + b_1 x$ 的斜率。

如果估计的回归方程的斜率为正（$b_1 > 0$），那么样本相关系数的符号为正；如果估计的回归方程的斜率为负（$b_1 < 0$），那么样本相关系数的符号为负。

对于 Armand 比萨饼连锁店的例子，对应于估计的回归方程 $\hat{y} = 60 + 5x$，判定系数的数值是 0.902 7。因为估计的回归方程的斜率是正的，所以由式（14-13），我们得到的样本相关系数是 $+\sqrt{0.902\,7} = +0.950\,1$。由于样本相关系数 $r_{xy} = +0.950\,1$，所以我们可以得出结论：x 和 y 之间存在一个强的正向线性关系。

在两变量之间存在一个线性关系的情况下，判定系数和样本相关系数都给出了它们之间线性关系强度的度量。判定系数给出的测度数值在 0 ~ 1，而样本相关系数给出的测度数值在 −1 ~ +1。虽然样本相关系数的适用范

围被限制在两变量之间存在线性关系的情况，但判定系数对非线性关系以及有两个或两个以上自变量的相关关系都适用。在这种意义上，判定系数有着更广泛的应用范围。

注释和评论

1. 在利用最小二乘法求出估计的回归方程和计算判定系数时，我们并没有对模型的误差项 ε 作出任何的概率假定，也没有对 x 和 y 之间关系的显著性进行统计检验。r^2 的数值比较大，只不过意味着最小二乘回归直线比较好地拟合了样本数据，也就是说，观测值比较紧密地围绕在最小二乘回归直线周围。但是仅仅利用 r^2，我们不能得出 x 和 y 之间的关系在统计上是否显著的结论。这样一个结论必须建立在进一步深入研究的基础上，包括对样本容量的研究和对最小二乘估计量适当的抽样分布性质的研究。
2. 在实际问题中，例如在社会科学问题中遇到的典型数据，尽管 r^2 的数值低于 0.25，通常也被认为是令人满意的。而在自然科学和生命科学问题中遇到的数据，经常会出现 r^2 的数值大于或等于 0.60。事实上，在某些情况下，我们有时还能遇到 r^2 的数值大于 0.90 的情形。在商务应用中，r^2 的数值将依据每一个应用的独特性质发生非常大的变化。

练习

方法

16. 第 2 题的数据如下。

x_i	3	12	6	20	14
y_i	55	40	55	10	15

这些数据的估计的回归方程是 $\hat{y} = 68 - 3x$

a. 计算 SSE，SST 和 SSR。

b. 计算判定系数 r^2。请对拟合优度作出评述。

c. 计算样本相关系数。

应用

☆18. 经《消费者报告》测试的六个立体声耳机的品牌、价格和整体评价得分的数据如下表所示（*Consumer Reports website*，2012 年 3 月 5 日）。整体评价得分是基于音质和降低环境噪音的效果而得到的。得分范围从 0（最低）到 100（最高）。根据这些数据得到的估计的回归方程是 $\hat{y} = 23.194 + 0.318x$，其中 x = 价格（美元）和 y = 整体得分。

品牌	价格（美元）	整体评价得分
Bose	180	76
Skullcandy	150	71
Koss	95	61
Phillips/O' Neill	70	56
Denon	70	40
JVC	35	26

a. 计算 SST，SSR 和 SSE。

b. 计算判定系数 r^2。请对拟合优度作出评述。

c. 样本相关系数的数值是多少？

20. *Bicycling* 是全球权威的自行车运动杂志，该杂志终年都对数以百计的自行车进行评论。其中，杂志的"公路赛车"栏目包含了对自行车的评论，而这些自行车的使用者主要是那些对自行车运动感兴趣的车手们。选择一辆比赛用自行车最重要的因素之一是车的重量。被该杂志评论的 10 辆公路比赛用自行车的重量（磅）和价格（美元）数据如下表所示（*Bicycling* website，2012 年 3 月 8 日）。

品牌	重量（磅）	价格（美元）
FELT F5	17.8	2 100
PINARELLO Paris	16.1	6 250
ORBEA Orca GDR	14.9	8 370
EDDY MERCKX EMX-7	15.9	6 200
BH RCI Ultegra	17.2	4 000
BH Ultralight 386	13.1	8 600
CERVELO S5 Team	16.2	6 000
GIANT TCR Advanced 2	17.1	2 580
WILIER TRIESTINA Gran Turismo	17.6	3 400
SPECIALIZED S-Works Amira SL4	14.1	8 000

a. 利用这些数据，建立一个估计的回归方程，使该方程在一辆自行车重量已知的情形下，能用来估计该车的价格。

b. 计算 r^2。估计的回归方程对这些观测数据的拟合好吗？

c. 若一辆自行车的重量是 15 磅，预测该车的价格。

22. 参阅第 5 题。下面的数据曾用于研究：在一般情形

下，较高价格的椭圆机是否具有较高的总体评价分数（《消费者报告》，2008 年 2 月）。

若 x = 价格（美元），y = 总体评价分，则估计的回归方程是 $\hat{y} = 58.158 + 0.008\,449x$。对于这些数据，SSE = 173.88。

a. 计算判定系数 r^2。

b. 估计的回归方程对这些数据的拟合好吗？请作出解释。

c. 样本相关系数的数值是多少？样本相关系数反映出在价格和总体评价分数之间存在强的还是弱的相关关系？

品牌和型号	价格（美元）	总体评价分数
Precor 5.31	3 700	87
Key Fitness CG2	2 500	84
Octane Fitness Q37e	2 800	82
LifeFitness X1 Basic	1 900	74
Nordic Track AudioStrider 990	1 000	73
Schwinn 430	800	69
Vision Fitness X6100	1 700	68
ProForm XP 520 Razor	600	55

14.4 模型的假定

在进行回归分析时，我们首先要对描述应变量和自变量之间关系的模型作出一些假定。对于简单线性回归情况，假定回归模型是

$$y = \beta_0 + \beta_1 x + \varepsilon$$

然后利用最小二乘法，分别求出模型参数 β_0 和 β_1 的估计值 b_0 和 b_1。得到估计的回归方程是

$$\hat{y} = b_0 + b_1 x$$

我们已经知道，判定系数 r^2 的值是估计的回归方程拟合优度的度量。然而，尽管 r^2 有一个较大的数值，但是在对假定模型的合理性作出进一步的分析完成之前，我们还不能应用这个估计的回归方程。确定假定的模型是否合理的一个重要步骤，是要对变量之间关系的显著性进行检验。回归分析中的显著性检验是以对误差项 ε 的下列假定为依据进行的。

关于回归模型 $y = \beta_0 + \beta_1 x + \varepsilon$ 的误差项 ε 的假定

1. 误差项 ε 是一个平均值或期望值为零的随机变量，即 $E(\varepsilon) = 0$。

这就意味着：因为 β_0 和 β_1 都是常数，所以有 $E(\beta_0) = \beta_0$ 和 $E(\beta_1) = \beta_1$；于是对于一个给定的 x 值，y 的期望值是

$$E(y) = \beta_0 + \beta_1 x \tag{14-14}$$

正如我们在前面已经指出的那样，称式（14-14）为回归方程。

2. 对所有的 x 值，ε 的方差都是相同的。我们用 σ^2 表示 ε 的方差。

这就意味着：y 关于回归直线的方差等于 σ^2，也就是说对所有的 x 值，y 的方差都是相等的。

3. ε 的值是相互独立的。

这就意味着：对于一个特定的 x 值，它所对应的 ε 值与任何其他的 x 值所对应的 ε 值不相关。于是，对于一个特定的 x 值，它所对应的 y 值与任何其他的 x 值所对应的 y 值也不相关。

4. 对所有的 x 值，误差项 ε 是一个正态分布的随机变量。

这就意味着：因为 y 是 ε 的一个线性函数，所以对所有的 x 值，y 也是一个正态分布的随机变量。

图 14-6 是对模型假定及其含义的一个说明。值得注意的是，从这个图形解释中可以看出，$E(y)$ 的值是根据被考虑的具体的 x 值的变化而变化。然而，不论 x 的值怎样变化，ε 和 y 的概率分布都是正态分布，并且具有相同的方差。在任一特定点上，误差项 ε 的具体数值依赖于 y 的实际值是大于还是小于 $E(y)$。

现在，我们必须记住，我们还曾经对变量 x 和 y 之间关系的形式做过一个假定或假设，即我们假定，两变量 x 和 y 之间关系的主要部分是一条表示为 $\beta_0 + \beta_1 x$ 的直线。我们还不能忽略这样一个事实，一些其他形式的模型，例如 $y = \beta_0 + \beta_1 x^2 + \varepsilon$，或许可能是描述两变量 x 和 y 之间相关关系的一个更好的模型。

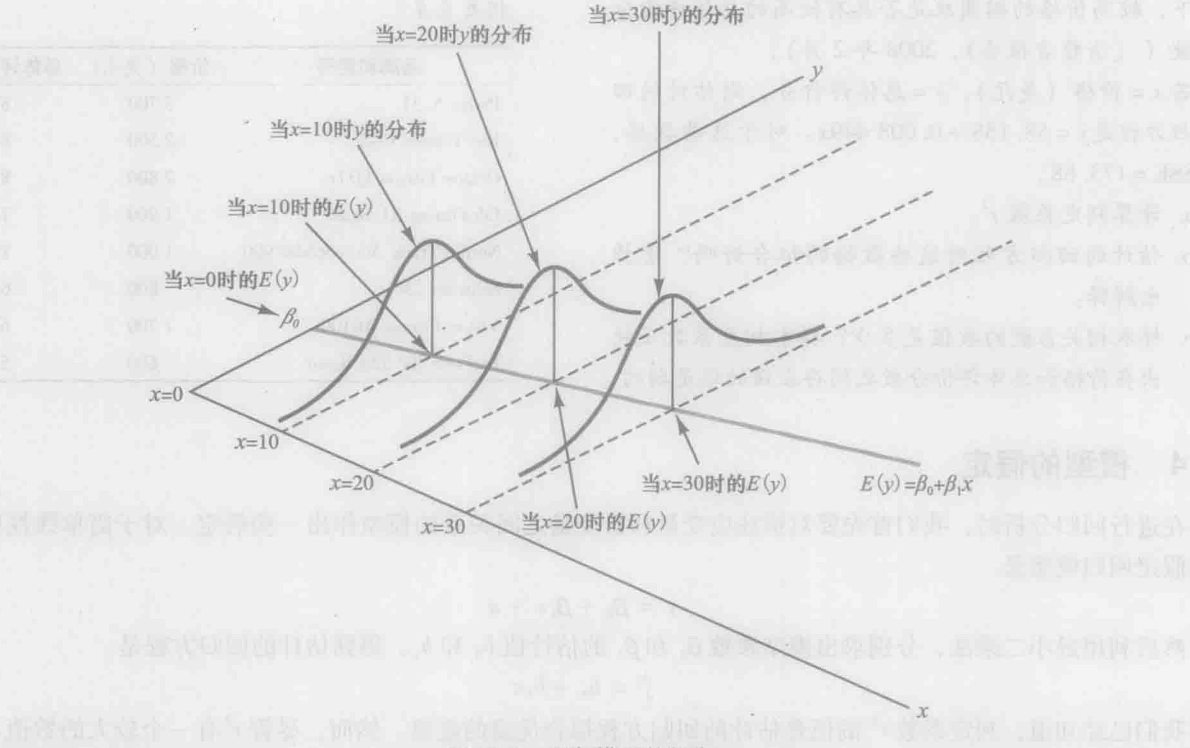


图 14-6 回归模型的假定

注：对于每一个 x 值，y 的分布都有相同的形状。

14.5 显著性检验

在一个简单线性回归方程里，y 的平均值或期望值是 x 的一个线性函数：$E(y)=\beta_0+\beta_1x$。如果 β_1 的值是零，$E(y)=\beta_0+0x=\beta_0$。在这种情况下，y 的平均值或期望值不依赖于 x 的值，因此我们的结论是，在两变量 x 和 y 之间不存在线性关系。换一种说法，如果 β_1 的值不等于零，我们的结论是，在两变量 x 和 y 之间存在线性关系。于是，为了检验两变量之间是否存在一个显著的回归关系，我们必须进行一个假设检验，用来判定 β_1 的值是否等于零。通常使用的检验方法有两种，这两种方法都需要知道回归模型误差项 ε 的方差 σ^2 的估计值。

14.5.1 σ^2的估计

从回归模型和它的假定中，我们可以得出结论：ε 的方差 σ^2 也是应变量 y 的值关于回归直线的方差。回想一下，我们曾经将 y 的值关于估计的回归直线的离差称为残差。所以，残差平方和 SSE 是实际观测值关于估计的回归直线变异性的度量。我们用 SSE 除以它自己的自由度，得到**均方误差**（mean square error，MSE）。均方误差给出了 σ^2 的一个估计量。

因为 $\hat{y}_i=b_0+b_1x_i$，所以 SSE 可以写成

$$\text{SSE}=\sum(y_i-\hat{y}_i)^2=\sum(y_i-b_0-b_1x_i)^2$$

每一个平方和都有一个与之相联系的数，这个数叫作自由度。统计学家已经证明，为了计算 SSE，必须估计两个参数（β_0 和 β_1），所以 SSE 的自由度是 $n-2$。于是，用 SSE 除以 $n-2$，就能计算出均方误差。MSE 是 σ^2 的一个无偏估计量。因为 MSE 的值给出了 σ^2 的一个估计，我们用记号 s^2 表示。

均方误差（σ^2的估计量）

$$s^2=\text{MSE}=\frac{\text{SSE}}{n-2} \tag{14-15}$$

在第 14.3 节 Armand 比萨饼连锁店的例子中，我们已经得到 SSE = 1 530，于是 σ^2 的一个无偏估计值是

$$s^2 = \text{MSE}\frac{1\,530}{8} = 191.25$$

为了估计 σ，我们计算 s^2 的平方根，所得到的结果 s 被称为**估计的标准误差**（standard error of the estimate）。

估计的标准误差

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}} \tag{14-16}$$

对于 Armand 比萨饼连锁店的例子，$s = \sqrt{\text{MSE}} = \sqrt{191.25} = 13.829$。在下面的讨论中，我们利用估计的标准误差，对两变量 x 和 y 之间的关系进行显著性检验。

14.5.2　t 检验

对于简单线性回归模型 $y = \beta_0 + \beta_1 x + \varepsilon$，如果 x 和 y 之间存在一个线性关系，我们必须有 $\beta_1 \neq 0$。t 检验的目的是我们能否断定 $\beta_1 \neq 0$。我们将利用样本数据去检验下面的关于参数 β_1 的假设。

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

如果 H_0 被拒绝，我们将会得到 $\beta_1 \neq 0$ 的结论，于是在两变量 x 和 y 之间存在一个统计上是显著的关系；如果 H_0 没有被拒绝，我们将没有充分的理由来断定，在两变量 x 和 y 之间存在一个统计上是显著的关系。β_1 的最小二乘估计量 b_1 的抽样分布性质给出了假设检验的基础。

首先让我们考虑一下，如果在同样的回归研究中，我们使用了不同的随机样本，将会出现什么情况。例如，在 Armand 比萨饼连锁店的例子中，假设我们使用了由另外 10 家连锁店组成的样本得到的季度销售收入数据。对这一新的样本进行回归分析，得到的估计的回归方程可能与我们前面得到的估计的回归方程 $\hat{y} = 60 + 5x$ 相类似，但是不可能得到完全相同的回归方程（截距恰好是 60，斜率恰好是 5）。实际上，最小二乘估计量 b_0 和 b_1 是样本统计量，它们有着自己的抽样分布。b_1 抽样分布的性质如下。

b_1 的抽样分布

期望值　$E(b_1) = \beta_1$

标准差　$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}} \tag{14-17}$$

分布形式　正态分布

注意，b_1 的期望值等于 β_1，所以 b_1 是 β_1 的无偏估计量。

因为 σ 的值未知，为了求出 σ_{b_1} 的一个估计 s_{b_1}，我们用 σ 的估计 s 代入式（14-17），于是我们得到下面 σ_{b_1} 的估计。

b_1 的估计的标准差[⊖]

$$s_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} \tag{14-18}$$

对于 Armand 比萨饼连锁店的例子，$s = 13.829$。于是，利用表 14-2 给出的结果，$\sum(x_i - \bar{x})^2 = 568$，我们得到 b_1 的估计的标准差为

⊖ b_1 的标准差 σ_{b_1} 也被称为 b_1 的标准误差。于是，s_{b_1} 给出了 b_1 的标准误差的一个估计。

$$s_{b_1} = \frac{13.829}{\sqrt{568}} = 0.5803$$

对于两个变量之间是否存在一个显著性关系的 t 检验的依据是，检验统计量

$$\frac{b_1 - \beta_1}{s_{b_1}}$$

是一个服从自由度为 $n=2$ 的 t 分布。如果原假设成立，则 $\beta_1=0$，并且 $t=b_1/s_{b_1}$。

现在让我们在 $\alpha=0.01$ 的显著性水平下，对 Armand 比萨饼连锁店的例子进行显著性检验。检验的统计量为

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{0.5803} = 8.62$$

从 t 分布表（表 D-2）中我们得到，当自由度为 $n-2=10-2=8$ 时，$t=3.355$ 的 t 分布上侧的面积是 0.005。于是对应于检验统计量 $t=8.62$ 的 t 分布上侧的面积一定小于 0.005。因为这是双侧检验，我们可以断定，与 $t=8.62$ 相联系的 p-值一定小于 2 倍的 0.005，即一定小于 $2\times0.005=0.01$。Excel 或 Minitab 给出的 p-值=0.000。因为 p-值小于 $\alpha=0.01$，所以我们拒绝 H_0，并且得到结论：β_1 显著不等于零。这一结果足以使我们断定，学生人数和季度销售收入之间存在一个显著的关系。[⊖]对于简单线性回归情形，总结显著性 t 检验的步骤如下。

简单线性回归显著性的 t 检验

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

检验的统计量

$$t = \frac{b_1}{s_{b_1}} \tag{14-19}$$

拒绝法则

p-值法：如果 p-值$\leqslant\alpha$，则拒绝 H_0

临界值法：如果 $t\leqslant -t_{\alpha/2}$ 或者 $t\geqslant t_{\alpha/2}$，则拒绝 H_0

式中，$t_{\alpha/2}$ 自由度为 $n-2$ 时，使 t 分布的上侧面积为 $\alpha/2$ 的 t 值。

14.5.3 β_1 的置信区间

β_1 的置信区间的公式是：

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

式中，b_1 为 β_1 的点估计量；$t_{\alpha/2}s_{b_1}$ 为边际误差。与这个置信区间相联系的置信系数是 $1-\alpha$；$t_{\alpha/2}$ 为自由度为 $n-2$时，使 t 分布的上侧面积为 $\alpha/2$ 的 t 值。

例如，对于 Armand 比萨饼连锁店的例子，假设我们希望建立一个 β_1 的 99% 的置信区间估计。对于置信系数 $\alpha=0.01$ 和自由度 $n-2-10-2=8$，我们从表 B-2 中，得到 $t_{0.005}=3.355$。于是，β_1 的 99% 的置信区间是

$$b_1 \pm t_{\alpha/2} s_{b_1} = 5 \pm 3.355 \times 0.5803 = 5 \pm 1.95$$

即为 3.05 ~ 6.95。

在对 β_1 的显著性进行 t 检验时，被检验的假设是

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

作为可供选择的一种方法，对于 Armand 比萨饼连锁店的数据，在 $\alpha=0.01$ 的显著性水平下，我们可以使用 99%

⊖ 本书配套光盘中的附录 14C 和附录 14D 说明了如何使用 Minitab 和 Excel 计算 p-值。

的置信区间得到假设检验的结论。因为β_1的假设值是0，而0没有包括在置信区间（3.05，6.95）里，所以我们可以拒绝H_0，并且可以断定，在学生人数和季度销售收入之间存在一个在统计上是显著的关系。在一般情况下，我们可以利用置信区间对β_1的任何双侧假设进行检验。如果β_1的假设值包括在置信区间里，则不拒绝H_0；否则，拒绝H_0。

14.5.4 *F*检验

建立在*F*概率分布基础上的*F*检验，也可以用来对回归方程进行显著性检验。在仅有一个自变量的情况时，*F*检验将得出与*t*检验同样的结论；即如果*t*检验表明$\beta_1 \neq 0$，因此变量之间存在一个显著的关系，那么，*F*检验也将表明变量之间存在一个显著的关系。但是，如果回归方程有两个或两个以上自变量的情况时，*F*检验仅仅能被用来检验回归方程总体的显著关系。

为了确定变量间的回归关系在统计上是否显著，我们所使用的*F*检验的基本原理是基于建立σ^2的两个独立的估计量。我们已经解释了，MSE是σ^2的一个估计量。如果原假设H_0：$\beta_1=0$成立，那么用回归平方和SSR除以它的自由度就给出了σ^2的另一个独立的估计量。这个估计量被称为来自于回归的均方，简称均方回归，用MSR表示。在一般情况下，

$$\text{MSR}=\frac{\text{SSR}}{\text{回归自由度}}$$

对在本书中我们所研究的模型，回归自由度总是等于模型中自变量的个数。于是

$$\text{MSR}=\frac{\text{SSR}}{\text{自变量的个数}} \tag{14-20}$$

由于在本章中，我们考虑的回归模型仅含有一个自变量，所以我们有MSR = SSR/1 = SSR。因此对Armand比萨饼连锁店的例子，MSR = SSR = 14 200。

如果原假设H_0：$\beta_1=0$成立，则MSR和MSE是σ^2的两个独立的估计量，并且MSR/MSE的抽样分布是服从分子的自由度为1，分母的自由度为$n-2$的*F*分布。所以当$\beta_1=0$时，MSR/MSE的值应接近于1。但是，如果原假设H_0：$\beta_1 \neq 0$不成立，MSR将高估σ^2，并且MSR/MSE的值将变得无穷大；于是，较大的MSR/MSE的值将导致拒绝H_0，并且我们可以断定，两变量x和y之间的关系在统计上是显著的。

现在我们对Armand比萨饼连锁店的例子进行*F*检验。检验的统计量是

$$F=\frac{\text{MSR}}{\text{MSE}}=\frac{14\,200}{191.25}=74.25$$

从表B-4中，我们查出，当分子的自由度为1，分母的自由度为$n-2=10-2=8$时，$F=11.26$的*F*分布上侧的面积是0.01。于是对应于检验统计量$F=74.25$的*F*分布上侧的面积一定小于0.01，因此我们可以断定，*p*-值一定小于0.01。Minitab或Excel都显示出*p*-值=0.000。因为*p*-值小于$\alpha=0.01$，所以我们拒绝H_0，并且可以断定，在学生人数和季度销售收入之间存在一个显著的关系[㊀]。对于简单线性回归情形，概括显著性*F*检验的步骤如下。

简单线性回归显著性的*F*检验

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

检验的统计量

$$F=\frac{\text{MSR}}{\text{MSE}} \tag{14-21}$$

拒绝法则

p-值法：如果*p*-值$\leqslant \alpha$，则拒绝H_0

临界值法：如果$F \geqslant F_\alpha$，则拒绝H_0

式中，F_α是分子自由度为1，分母自由度为$n=2$时，使*F*分布的上侧面积为α的*F*值。[㊁]

㊀ 对简单线性回归情形，*F*检验和*t*检验给出了同样的结果。

㊁ 如果H_0不成立，MSE仍然是σ^2的一个无偏估计量，而MSR高估σ^2。如果H_0成立，MSE和MSR都是σ^2的无偏估计量，在这种情况下，MSR/MSE的值应接近于1。

在第13章中，我们已经详细讨论了方差分析（ANOVA）问题，并且指出了如何使用一个ANOVA表简练地概括方差分析的运算过程。一个类似的ANOVA表能用来概括回归显著性的F检验的计算结果。表14-5是简单线性回归ANOVA表的一般形式。表14-6是对Armand比萨饼连锁店例子进行F检验计算过程的ANOVA表。第1列中的回归、误差和总和是方差的三个来源；第2列中的SSR、SSE和SST是对应的三个平方和；第3列是这些平方和的自由度，SSR为1、SSE为$n-2$和SST为$n-1$；第4列是两个均方MSR和MSE的值；第5列是统计量$F=$ MSR/MSE的值，第6列是对应于第5列的F值的p-值。几乎所有回归分析的计算机打印输出都包含了一张ANOVA表，而这张ANOVA表概括了显著性F检验的步骤[⊖]。

表14-5 简单线性回归ANOVA表的一般形式

方差来源	平方和	自由度	均方	F	p-值
回归	SSR	1	$MSR=\frac{SSR}{1}$	$F=\frac{MSR}{MSE}$	
误差	SSE	$n-2$	$MSE=\frac{SSE}{n-2}$		
总和	SST	$n-1$			

表14-6 Armand比萨饼连锁店例子的ANOVA表

方差来源	平方和	自由度	均方	F	p-值
回归	14 200	1	$\frac{14\,200}{1}=14\,200$	$\frac{14\,200}{191.25}=74.25$	0.000
误差	1 530	8	$\frac{1\,530}{8}=191.25$		
总和	15 730	9			

14.5.5 关于显著性检验解释的几点注意

拒绝原假设H_0：$\beta_1=0$并且作出变量x和y之间存在显著性关系的结论，并不意味着我们能作出变量x和y之间存在一个因果关系的结论。只有当分析人员有着理论上的充分证据，能够证明变量之间确实存在因果关系，我们才能认为变量之间存在这样一个因果关系的结论是合理的。在Armand比萨饼连锁店例子中，我们能作出在学生人数x和季度销售收入y之间存在一个显著关系的结论，而且，估计的回归方程$\hat{y}=60+5x$给出了这一显著关系的最小二乘估计结果。但是我们不能断定学生人数x的变化引起了季度销售收入y的变化，因为我们只不过识别了一个统计上显著的关系。这样一个因果关系结论的合理性，一方面要从理论上给予证实，另一方面还要依靠分析人员出色的判断能力。Armand比萨饼连锁店的管理人员觉得，学生人数的增长很可能是季度销售收入增长的一个原因。于是，显著性检验的结果能够使他们作出在学生人数和季度销售收入之间确实存在一个因果关系的结论[⊖]。

另外，由于我们只不过是拒绝了H_0：$\beta_1=0$和证实了变量x和y之间存在统计显著性关系，但这并不能让我们作出变量x和y之间存在线性关系的结论。我们仅仅能说明在x的样本观测值范围以内，x和y是相关的，而且这个线性关系只是在x的样本观测值范围里，解释了y的变异性的显著部分。图14-7说明了这种情形。显著性检验的结果要求我们否定原假设H_0：$\beta_1=0$，并且作出了x和y之间存在显著关系的结论，但是图14-7表明，在x和y之间的关系实际上不是线性关系。虽然在x的样本观测值范围以内，估计的回归方程$\hat{y}=b_0+b_1x$给出了变量x和y之间关系一个很好的线性近似，但对这个范围以外的x值却很差。

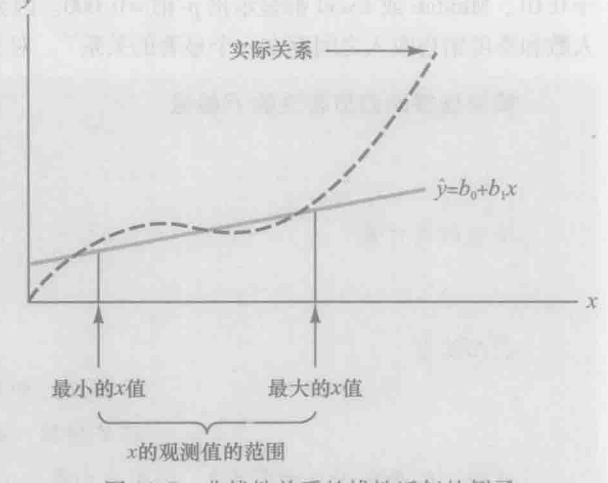


图14-7 非线性关系的线性近似的例子

假设变量x和y之间存在一个显著的关系，利用估计的回归方程，对于x的样本观测值范围以内的x值进行预测，我们应该是完全有把握的。对于Armand比萨饼连锁店的例子，x的样本观测值的取值范围为2~26。除非有理由相信，

⊖ 在每一个方差分析表中，总平方和是回归平方和与误差平方和之和；同时，总平方和的自由度是回归平方和的自由度与误差平方和的自由度之和。

⊖ 回归分析能用来识别变量之间是如何相互联系的，而不能用来作为变量之间存在一个因果关系的根据。

超出这个范围模型仍是适宜的，在一般情形下，在自变量 x 的取值范围以外进行预测应十分小心谨慎。对于Armand比萨饼连锁店的例子，因为我们已经知道，在0.01显著性水平下，回归关系是显著的，所以对于有2 000～26 000名学生的校园附近的连锁店，利用这个显著的回归关系来预测连锁店季度销售收入应该是完全有把握的。

注释和评论

1. 由于我们在第14.4节中对误差项作出了一些假定，才使我们在本节中进行统计显著性检验成为可能。b_1 抽样分布的性质以及随后的 t 检验和 F 检验都是从这些假设直接得出的。
2. 不能把统计显著性与实际显著性搞混淆了。当样本容量非常大时，对于小的 b_1 的值我们也能得到在统计上是显著的结果。在这种情形下，要作出在实际中存在一个显著性关系的结论，必须特别小心。
3. 变量 x 和 y 之间存在一个线性关系的显著性检验，也可以利用样本相关系数 r_{xy} 来完成。用 ρ_{xy} 表示总体相关系数，提出的假设如下。

$$H_0:\rho_{xy}=0$$
$$H_a:\rho_{xy}\neq 0$$

如果 H_0 被拒绝，那么就能作出变量 x 和 y 之间存在一个显著性关系的结论。有关这一检验的详细阐述在附录14B中给出。不过，利用相关系数进行显著性检验与本节前面介绍的 t 检验和 F 检验给出了同样的结果。所以，如果已经进行了 t 检验或 F 检验，那么就不必再利用相关系数进行显著性检验了。

练习

方法

24. 第2题的数据如下。

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a. 利用式（14-15）计算均方误差。

b. 利用式（14-16）计算估计的标准误差。

c. 利用式（14-18）计算 b_1 的估计的标准差。

d. 利用 t 检验，在 $\alpha=0.05$ 的显著性水平下，检验下面的假设：

$$H_0:\beta_1=0$$
$$H_a:\beta_1\neq 0$$

e. 利用 F 检验，在 $\alpha=0.05$ 的显著性水平下，检验（d）中的假设。请用ANOVA表的格式表示所得到的结果。

应用

☆26. 在第18题中，经《消费者报告》测试的6个立体声耳机的价格（美元）和整体评价得分的数据如下表所示（*Consumer Reports* website，2012年3月5日）。

品牌	价格（美元）	整体评价得分	品牌	价格（美元）	整体评价得分
Bose	180	76	Phillips/O'Neill	70	56
Skullcandy	150	71	Denon	70	40
Koss	95	61	JVC	35	26

a. 在 $\alpha=0.05$ 的显著性水平下，t 检验是否表明在价格和整体评价得分之间存在一个显著的关系？你的结论是什么？

b. 在 $\alpha=0.05$ 的显著性水平下，利用 F 检验，检验变量间的显著关系。你的结论是什么？

c. 对这些数据，作出ANOVA表。

28. 对第8题的等级评估数据，若设 x＝执行速度的能力，y＝电子交易的整体满意度，得到的估计的回归方程是 $\hat{y}=0.2046+0.9077x$。在 $\alpha=0.05$ 的显著性水平下，检验执行速度与整体满意度是否相关。作出ANONA表。你的结论是什么？

30. 参阅第5题，下面的数据曾用于来研究：在一般情形下，较高价格的椭圆机是否具有较高的总体评价分数（《消费者报告》，2008年2月）。若设 x＝价格（美元），y＝总体评价分，得到的估计的回归方程是 $\hat{y}=58.158+0.008449x$。对于这些数据，SSE＝173.88，SST＝756。是否有证据表明椭圆机的价格与总体评价分数这两个变量之间存在一个显著的关系？

品牌和型号	价格（美元）	总体评价分数
Precor 5.31	3 700	87
Key Fitness CG2	2 500	84
Octane Fitness Q37e	2 800	82
LifeFitness X1 Basic	1 900	74

（续）

品牌和型号	价格（美元）	总体评价分数
Nordic Track AudioStrider 990	1 000	73
Schwinn 430	800	69
Vision Fitness X6100	1 700	68
ProForm XP 520 Razor	600	55

31. 对于第 20 题中的 10 辆公路比赛用自行车的数据，若设 x = 重量（磅），y = 价格（美元），得到的估计的回归方程是 $\hat{y} = 28\,574 - 1\,439x$。（*Bicycling* website，2012 年 3 月 8 日）对于这些数据，SSE = 7 102 922.54，SST = 52 120 800。利用 F 检验，在 $\beta = 0.05$ 的显著性水平下，确定自行车的重量和价格是否相关。

14.6 应用估计的回归方程进行估计和预测

简单线性回归模型是我们对 x 和 y 之间关系所做的一个假定。然后，我们利用最小二乘法，得到了估计的简单线性回归方程。如果在变量 x 和 y 之间存在一个显著的关系，并且判定系数表明估计的回归方程对样本数据的拟合是一个好的拟合，那么我们就能利用估计的回归方程进行估计和预测。

在 Armand 比萨饼连锁店的例子中，估计的回归方程是 $\hat{y} = 60 + 5x$。在第 14.1 节的末尾，我们曾经说过，对于 x 的一个特定值，$\hat{y}$ 可以被用来作为 y 的平均值或期望值 $E(y)$ 的一个点估计，或者 $\hat{y}$ 也可以被用来作为 y 的一个个别值的预测值。例如，在 Armand 比萨饼连锁店的例子中，假定对于所有位于有 10 000 名学生的校园附近的连锁店，管理人员希望得到平均季度销售收入一个点估计。利用估计的回归方程 $\hat{y} = 60 + 5x$，对于 $x = 10$（10 000 名学生），我们有 $\hat{y} = 60 + 5 \times 10 = 110$。于是，对于位于有 10 000 名学生的校园附近的所有连锁店，平均季度销售收入的点估计值是 110 000 美元。在这种情况下，当 $x = 10$ 时，我们用 $\hat{y}$ 作为 y 的平均值的点估计值。

对于 x 的一个特定值，我们也可以使用估计的回归方程来预测 y 的一个个别值。例如，在有 10 000 名学生的 Talbot 大学附近开了一家新的 Armand 比萨饼连锁店，管理人员希望预测该家连锁店的季度销售收入，于是我们计算 $\hat{y} = 60 + 5 \times 10 = 110$。因此，对这家新的连锁店，我们预测的季度销售收入 110 000 美元。在这种情况下，当 $x = 10$ 时，我们用 $\hat{y}$ 作为 y 的一个新观测值的预测值。

当我们使用估计的回归方程来估计 y 的平均值或者来预测 y 的一个个别值时，显然估计或预测都依赖于 x 的一个给定值。因此，当我们在更深入讨论有关估计和预测问题时，下面的符号将有助于问题的阐述。

x^* 表示自变量 x 的一个给定值

y^* 表示当 $x = x^*$ 时，应变量 y 的可能值，它是一个随机变量

$E(y^*)$ 表示当 $x = x^*$ 时，应变量 y 的平均值或期望值

$\hat{y}^* = b_0 + b_1 x^*$ 表示当 $x = x^*$ 时，$E(y^*)$ 的点估计值和 y^* 的一个个别值的预测值

为了说明这些符号的使用，假设对于所有位于有 10 000 名学生的校园附近的 Armand 比萨饼连锁店，我们要估计它们的平均季度销售收入。在这种情况下，$x^* = 10$，对于所有的 $x^* = 10$ 的 Armand 比萨饼连锁店，$E(y^*)$ 表示未知的平均季度销售收入。于是，$E(y^*)$ 的点估计值是 $\hat{y}^* = 60 + 5 \times 10 = 110$，或 110 000 美元。但是，使用这些符号，$\hat{y}^* = 110$ 也是有 10 000 名学生的 Talbot 大学附近一家新开的 Armand 比萨饼连锁店的季度销售收入的预测值。

14.6.1 区间估计

点估计值和预测值不能提供有关估计和（或）预测精度的任何信息。因此我们必须建立置信区间和预测区间。**置信区间**（confidence interval），它是对于 x 的一个给定值，y 的平均值的一个区间估计。**预测区间**（prediction interval），它是对于 x 的一个给定值，对应 y 的一个新的观测值，也即 y 的一个个别值进行预测的一个区间估计。虽然对于 x 的一个给定值，y 的一个个别值的预测值和 y 的平均值的点估计值是相同的，但是这两种情况得到的区间估计是不同的。我们将要说明，预测区间的边际误差比较大。[⊖] 我们首先说明如何建立一个 y 的平均值的区间估计。

14.6.2 y 的平均值的置信区间

在一般情况下，我们不能期望估计值 $\hat{y}^*$ 恰好等于 $E(y^*)$。如果我们想要作出 $\hat{y}^*$ 是如何接近真实的平均值

⊖ 置信区间和预测区间都表明了回归结果的精度。比较窄的区间表明有一个比较高的精度。

$E(y^*)$的推断，我们将不得不估计$\hat{y}^*$的方差。我们用$s^2_{\hat{y}^*}$表示估计的$\hat{y}^*$的方差，它的计算公式是

$$s^2_{\hat{y}^*} = s^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \tag{14-22}$$

$\hat{y}^*$的估计的标准差由式（14-22）的平方根给出。

$$s_{\hat{y}^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \tag{14-23}$$

对 Armand 比萨饼连锁店的例子，在第 14.5 节中已给出了 $s = 13.829$ 的计算结果。因为 $x^* = 10$，$\bar{x} = 14$ 和 $\sum(x_i - \bar{x})^2 = 568$，我们利用式（14-23）得到

$$s_{\hat{y}^*} = 13.829\sqrt{\frac{1}{10} + \frac{(10-14)^2}{568}} = 13.829\sqrt{0.1282} = 4.95$$

置信区间的一般表达式如下。

$E(y^*)$ 的置信区间

$$\hat{y}^* \pm t_{\alpha/2}s_{\hat{y}^*} \tag{14-24}$$

式中，$1-\alpha$ 为置信系数；$t_{\alpha/2}$ 为自由度为 $n-2$ 时，使 t 分布的上侧面积为 $\alpha/2$ 的 t 值。[⊖]

对于所有位于有 10 000 名学生的校园附近的 Armand 比萨饼连锁店，我们利用式（14-24），建立一个置信水平为 95% 的平均季度销售收入 $E(y^*)$ 的置信区间估计。对 $\alpha/2 = 0.025$，自由度为 $n-2 = 10-2 = 8$，我们利用本书配套光盘中表 B-2，得到 $t_{0.025} = 2.306$。因为$\hat{y}^* = 110$，边际误差 $t_{\alpha/2}s_{\hat{y}^*} = 2.306 \times 4.95 = 11.415$，于是，置信水平为 95% 的置信区间估计是

$$110 \pm 11.415$$

若以美元为单位，对于所有位于有 10 000 名学生的校园附近的 Armand 比萨饼连锁店，置信水平为 95% 的平均季度销售收入的置信区间是 110 000 美元 ±11 415 美元。所以，当学生人数是 10 000 人时，置信水平为 95% 的平均季度销售收入的置信区间是（98 585 美元，121 415 美元）。

注意，当 $x^* - \bar{x} = 0$ 时，由式（14-23）给出的$\hat{y}^*$的估计的标准差最小。在这种情形下，$\hat{y}^*$的估计的标准差为

$$s_{\hat{y}^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n}}$$

这就是说，每当我们取 $x^* = \bar{x}$ 时，就能得到 y 的平均值最佳的或是最精确的估计量。事实上，x^*偏离 $\bar{x}$ 愈远，差 $x^* - \bar{x}$ 就变得愈大，结果是 x^*偏离 $\bar{x}$ 愈远，y 的平均值的置信区间就变得愈宽。这种情形如图 14-8 所示。

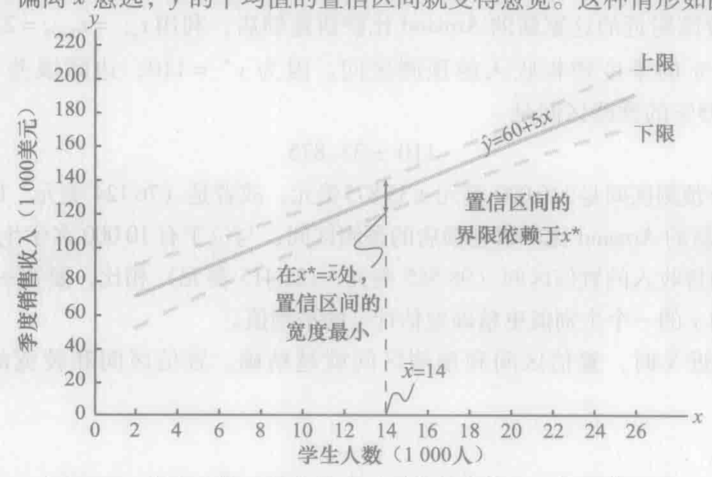


图 14-8　学生人数 x 的值给定时平均销售收入 y 的置信区间

⊖ 这个区间估计的边际误差是 $t_{\alpha/2}s_{\hat{y}^*}$。

14.6.3 y的一个个别值的预测区间

假设，我们不对位于有10 000名学生的校园附近的所有Armand比萨饼连锁店估计平均季度销售收入，我们只希望预测位于有10 000名学生的Talbot大学附近的一家新的Armand比萨饼连锁店的季度销售收入。正如前面所述，对应于自变量x的给定值x^*，应变量y的值y^*的预测值是$\hat{y}^* = b_0 + b_1 x^*$。对于位于Talbot大学附近的这家新的连锁店，我们有$x^* = 10$，对应的季度销售收入的预测值为$\hat{y}^* = 60 + 5 \times 10 = 110$或110 000美元。注意，对于位于Talbot大学附近的这家新的连锁店，该店的季度销售收入预测值与位于有10 000名学生的校园附近的所有Armand比萨饼连锁店的平均销售收入的点估计值是相同的。

为了建立预测区间，让我们首先确定，当$x = x^*$时，y的一个预测值$\hat{y}^*$的方差。这个方差由以下两部分组成。

1. y^*关于平均值$E(y^*)$的方差，它的估计量由s^2给出。
2. 利用$\hat{y}^*$估计$E(y^*)$的方差，它的估计量由$s^2_{\hat{y}^*}$给出。

当$x = x^*$时，应变量y的值y^*的预测值是$\hat{y}^*$，我们用s^2_{pred}表示y^*的预测值$\hat{y}^*$的估计的方差，它的计算公式是

$$s^2_{\text{pred}} = s^2 + s^2_{\hat{y}^*} = s^2 + s^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right] = s^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right] \tag{14-25}$$

因此，y^*的预测值$\hat{y}^*$的估计的标准差为

$$s_{\text{pred}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{14-26}$$

在Armand比萨饼连锁店的例子中，对于位于有10 000名学生的Talbot大学附近的这家新的连锁店，与季度销售收入预测值相对应的估计的标准差为

$$s_{\text{pred}} = 13.829\sqrt{1 + \frac{1}{10} + \frac{(10-14)^2}{568}} = 13.829\sqrt{1.128\,2} = 14.69$$

预测区间的一般表达式如下。

y^*的预测区间

$$\hat{y}^* \pm t_{\alpha/2} s_{\text{pred}} \tag{14-27}$$

式中，$1-\alpha$为置信系数；$t_{\alpha/2}$为自由度为$n-2$时，使t分布的上侧面积为$\alpha/2$的t值。[⊖]

对于位于Talbot大学校园附近的这家新的Armand比萨饼连锁店，利用$t_{\alpha/2} = t_{0.025} = 2.306$和$s_{\text{pred}} = 14.69$，我们能建立一个置信水平为95%的季度销售收入的预测区间。因为$y^* = 110$，边际误差$t_{\alpha/2} s_{\text{pred}} = 2.306 \times 14.69 = 33.875$，于是置信水平为95%的预测区间是

$$110 \pm 33.875$$

若以美元为单位，这个预测区间是110 000美元±33 875美元，或者是（76 124美元，143 875美元）。注意，位于Talbot大学校园附近的这家新的Armand比萨饼连锁店的预测区间，与位于有10 000名学生的校园附近的所有Armand比萨饼连锁店的平均季度销售收入的置信区间（98 585美元，121 415美元）相比，要宽一些。二者的区别说明了这样一个事实，我们能比预测y的一个个别值更精确地估计y的平均值。

当自变量的值x^*越接近$\bar{x}$时，置信区间和预测区间就越精确。置信区间和较宽的预测区间的一般形状如图14-9所示。[⊜]

⊖ 这个预测区间的边际误差是$t_{\alpha/2} s_{\text{pred}}$。

⊜ 在一般情况下，置信区间的界限和预测区间的界限都是弯曲的。

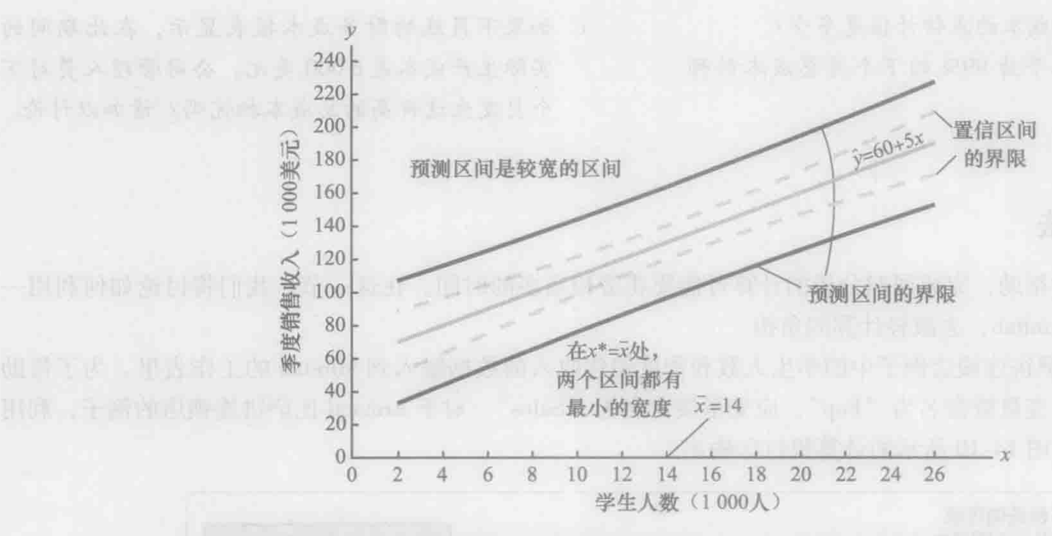


图14-9 学生人数 x 的值给定时销售收入 y 的置信区间和预测区间

注释和评论

对于一个新的观测值，预测区间用来预测应变量的值。作为一个例子，考虑位于有10 000名学生的Talbot大学附近的一家新的Armand比萨饼连锁店，我们说明如何对该连锁店的季度销售收入建立一个预测区间。实际上，对于表14-1 Armand比萨饼连锁店的样本数据，$x=10$ 并不是其中一个连锁店的学生人数，但这并不意味着不能对样本数据中的 x 值建立预测区间。但是对于组成表14-1数据的10家比萨饼连锁店，对其中任何一家建立季度销售收入的预测区间都是毫无意义的，因为我们已经知道了每一家连锁店的季度销售收入。换句话说，预测区间仅对新的事物才有意义，在这种情况下，是对应于一个 x 的特定值的一个新的观测值，而这个 x 的特定值可能等于或也可能不等于样本中的 x 值。

练 习

方法

☆32. 第1题的数据如下。

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. 当 $x=4$ 时，利用式（14-23）估计 $\hat{y}^*$ 的标准差。

b. 当 $x=4$ 时，利用式（14-24）建立一个 y 的期望值的置信水平为95%的置信区间。

c. 当 $x=4$ 时，利用式（14-26）估计 y 的一个个别值的标准差。

d. 当 $x=4$ 时，利用式（14-27）建立一个 y 的一个个别值的置信水平为95%的预测区间。

34. 第3题的数据如下。

x_i	2	6	9	13	20
y_i	7	18	9	26	23

当 $x=12$ 时，建立置信水平为95%的置信区间和预测区间。请作出解释：为什么这两个区间估计是不同的。

应用

36. 在第7题中，设 y = 对新客户账户的年销售额（1 000美元），x = 工龄（年），由10名售货员组成了一个随机样本数据，得到估计的回归方程是 $\hat{y}=80+4x$，$\bar{x}=7$，$\sum(x_i-\bar{x})^2=142$，$s=4.609\,8$。

a. 对于所有有9年工龄的售货员，建立一个置信水平为95%的平均年销售额的置信区间。

b. 公司正在考虑聘请有9年工龄的售货员Tom Smart，建立Tom Smart的一个置信水平为95%的年销售额的预测区间。

c. 讨论你在a和b中得到的答案的不同之处。

38. 参阅第21题，对于一个特定的制造业，利用产量和总成本的数据，得到估计的回归方程为 $\hat{y}=1\,246.67+7.6x$。

a. 公司的生产计划表明，下个月必须生产500件

产品。下个月总成本的点估计值是多少？

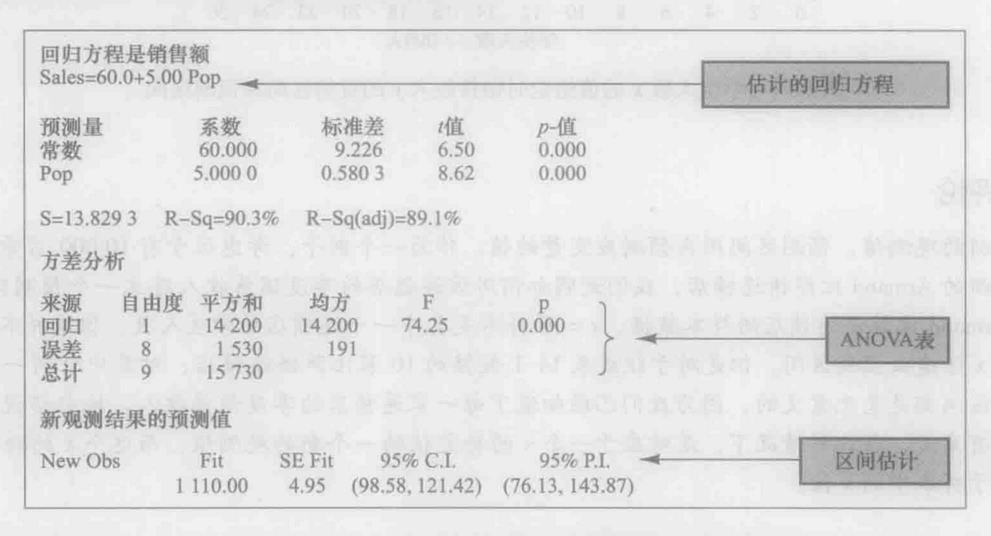
b. 建立一个置信水平为99%的下个月总成本的预测区间。

c. 如果下月底的财务成本报表显示，在此期间的实际生产成本是6 000美元。公司管理人员对下个月发生这样高的总成本担忧吗？请加以讨论。

14.7　计算机解法

如果没有计算机的帮助，完成回归分析的计算可能要花费相当多的时间。在这一节，我们将讨论如何利用一个计算机软件包，如Minitab，去减轻计算的负担。

我们将Armand比萨饼连锁店例子中的学生人数和季度销售收入的数据输入到Minitab的工作表里。为了帮助解释计算机的输出，自变量被命名为“Pop”，应变量被命名为“Sales”。对于Armand比萨饼连锁店的例子，利用Minitab，我们得到了如图14-10所示的计算机打印输出[⊖]。



回归方程是销售额
Sales=60.0+5.00 Pop

预测量	系数	标准差	t值	p-值
常数	60.000	9.226	6.50	0.000
Pop	5.000 0	0.580 3	8.62	0.000

S=13.829 3　R-Sq=90.3%　R-Sq(adj)=89.1%

方差分析

来源	自由度	平方和	均方	F	p
回归	1	14 200	14 200	74.25	0.000
误差	8	1 530	191		
总计	9	15 730			

新观测结果的预测值

New Obs	Fit	SE Fit	95% C.I.	95% P.I.
1	110.00	4.95	(98.58, 121.42)	(76.13, 143.87)

图14-10　Armand比萨饼连锁店问题的Minitab输出

1. Minitab打印出来的估计的回归方程为“Sales = 60.0 + 5.00 Pop”。

2. Minitab打印出来的表显示了系数b_0和b_1的值，每一个系数的标准差，每一个系数除以它的标准差得到的t值以及与t检验相联系的p-值。因为p-值为零（保留小数点至第3位），所以样本结果表明，应拒绝原假设（H_0：$\beta_1=0$）。另一种可供选择的检验方法是，我们将8.62（位于t值列）与一个适当的临界值进行比较。这种t检验的步骤我们已经在第14.5节中做过叙述。

3. Minitab打印出来的输出结果不但给出了估计的标准差“$s = 13.8293$”，而且给出了有关拟合优度的信息。注意，“R - sq = 90.3%”是用百分数表示的判定系数。数值“R - sq(adj) = 89.1%”将在第15章中讨论。

4. ANOVA表被打印在标题“Analysis of Variance”的下面。Minitab用横行标题“Residual Error”表示变异的误差来源。注意，“DF”是自由度的缩写，MSR为14 200，MSE为191。F值是这两个值之比，为74.25，对应的p-值为0.000。因为p-值为零（保留小数点至第3位），所以我们可以断定，“Sales”和“Pop”之间存在一个统计上显著的关系。

5. 期望销售收入的置信水平为95%的置信区间估计，以及在有10 000名学生的校园附近的一家个别的Armand比萨饼连锁店的销售收入的置信水平为95%的预测区间估计，打印在ANOVA表的下面。正如我们在第14.6节中已经给出的那样，置信区间是（98.58，121.42），预测区间是（76.13，143.87）。

⊖ 由Minitab生成的输出的必要步骤在本书所附光盘附录14C中给出。

练 习

方法

☆40. 若 x 表示公寓住宅的年租金（1 000 美元），y 表示该公寓住宅的销售价格（1 000 美元），一家房地产公司的营销部门对 x 和 y 之间的关系进行回归分析。采集了近期出售的一些公寓住宅的数据，得到的计算机输出如下。

回归方程是 $\hat{y}=20.0+7.21x$

预测量	系数	标准差	t 值
常数	20.000	3.221 3	6.21
x	7.210	1.362 6	5.29

方差分析

来源	自由度	平方和
回归	1	41 587.3
误差	7	
总计	8	51 984.1

a. 样本中有多少幢公寓住宅？

b. 写出估计的回归方程。

c. s_{b_1} 的值是多少？

d. 利用 F 统计量，在 $\alpha=0.05$ 的显著性水平下，检验 x 和 y 之间关系的显著性。

e. 对年租金为 50 000 美元的一幢公寓住宅，估计它的销售价格。

42. 若回归模型的 x 表示在一家分店中工作的售货员的人数，y 表示这家分店的年销售额（1 000 美元），对有关数据进行回归分析的计算机输出如下。

a. 写出估计的回归方程。

b. 对这一问题进行回归研究时涉及多少家分店？

c. 计算 F 统计量，在 $\alpha=0.05$ 的显著性水平下，检验 x 与 y 之间关系的显著性。

d. 若孟菲斯分店有 12 名售货员，预测该分店的年销售收入。

回归方程是 $\hat{y}=80.0+50.00x$

预测量	系数	标准差	t 值
常数	80.0	11.333	7.06
x	50.0	5.482	9.12

方差分析

来源	自由度	平方和	均方
回归	1	6 828.6	6 828.6
误差	28	2 298.8	82.1
总计	29	9 127.4	

44. 汽车大赛、优质驾校以及由汽车俱乐部经营的司机培训计划受人们喜爱的程度持续增长。所有这些活动的参加者都被要求必须佩戴经 Snell 纪念基金会认证的头盔。Snell 纪念基金会是一个非营利性的组织，该基金会致力于头盔安全标准的研究、培训、测试和新产品的开发。Snell 纪念基金会的额定专业头盔“SA”（Sports Application，体育运动专用）是专门为赛车运动设计的，这种头盔具有极强的耐冲击性和高防火性能。在选择头盔时，一个关键的因素是头盔的重量，因为较轻的头盔往往对颈部的压力较小。下表是 18 个 SA 头盔的重量和价格的统计数据（SoloRacer website，2008 年 4 月 20 日）。

a. 以重量为自变量，绘制出这些数据的散点图。

b. 在头盔的重量和价格这两个变量之间显现出什么关系吗？

c. 求出估计的回归方程，使这个方程在头盔的重量已知时，能用来预测头盔的价格。

d. 在 $\alpha=0.05$ 的显著性水平下，检验这两个变量之间关系的显著性。

e. 估计的回归方程对观测数据的拟合好吗？请作出解释。

头盔	重量（盎司）	价格（美元）
Pyrotect Pro Airflow	64	248
Pyrotect Pro Airflow Graphics	64	278
RCi Full Face	64	200
RaceQuip RidgeLine	64	200
HJC AR-10	58	300
HJC Si-12	47	700
HJC HX-10	49	900
Impact Racing Super Sport	59	340
Zamp FSA-1	66	199
Zamp RZ-2	58	299
Zamp RZ-2 Ferrari	58	299
Zamp RZ-3 Sport	52	479
Zamp RZ-3 Sport Painted	52	479
Bell M2	63	369
Bell M4	62	369
Bell M4 Pro	54	559
G Force Pro Force 1	63	250
G Force Pro Force 1 Grafx	63	280

14.8 残差分析：证实模型假定

正如我们前面提到的，第 i 次观测的残差是应变量的观测值（y_i）与它的预测值（$\hat{y}_i$）之差。[⊖]

第 i 次观测的残差

$$y_i - \hat{y}_i \tag{14-28}$$

式中，y_i 代表应变量的观测值；$\hat{y}_i$ 代表应变量的预测值。

换言之，第 i 次观测的残差是利用估计的回归方程去预测应变量的值 y_i 而产生的误差。对 Armand 比萨饼连锁店的例子，残差的计算列示在表 14-7 中。应变量的观测值在第二列，利用估计的回归方程 $\hat{y}=60+5x$ 得到应变量的预测值在第三列，对应的残差在第四列。对这些残差进行分析将帮助我们确定，关于回归模型作出的那些假定是否成立。

表 14-7 Armand 比萨饼连锁店例子的残差

学生人数 x_i	销售收入 y_i	估计的销售收入 $\hat{y}_i=60+5x_i$	残差 $y_i-\hat{y}_i$	学生人数 x_i	销售收入 y_i	估计的销售收入 $\hat{y}_i=60+5x_i$	残差 $y_i-\hat{y}_i$
2	58	70	−12	16	137	140	−3
6	105	90	15	20	157	160	−3
8	88	100	−12	20	169	160	9
8	118	100	18	22	149	170	−21
12	117	120	−3	26	202	190	12

现在让我们重温一下关于 Armand 比萨饼连锁店例子的回归假定。假设简单线性回归模型为

$$y=\beta_0+\beta_1x+\varepsilon \tag{14-29}$$

这个模型表示，我们假定季度销售收入（y）是学生人数（x）的一个线性函数加上一个误差项 ε。在第 14.4 节中，我们对误差项 ε 作出了以下假定。

1. $E(\varepsilon)=0$。
2. 对所有的 x 值，ε 的方差都是相同的，用 σ^2 表示。
3. ε 的值是相互独立的。
4. 误差项 ε 服从正态分布。

这些假定对于利用 t 检验和 F 检验来确定 x 和 y 之间的关系是否显著，以及对于在第 14.6 节中所介绍的置信区间估计和预测区间估计，都提供了理论上的依据。如果关于误差项 ε 的假定显得不那么可靠，那么有关回归关系的显著性假设检验和区间估计的结果可能会站不住脚。

残差提供了有关误差项 ε 的最重要的信息，因此，残差分析是确定误差项 ε 的假定是否成立的重要步骤。许多残差分析都是在对残差图形的仔细考察基础上完成的。在本节，我们将讨论以下几种残差图。

1. 关于自变量 x 的值的残差图。
2. 关于应变量的预测值 $\hat{y}$ 的残差图。
3. 标准化残差图。
4. 正态概率图。

14.8.1 关于 x 的残差图

关于自变量 x 的**残差图**（residual plot）的做法是，用水平轴表示自变量的值，用纵轴表示对应的残差值。每个残差都用图上的一个点来表示。对于每一个点，第一个坐标由 x_i 的值给出，第二个坐标由对应的第 i 个残差

⊖ **残差分析**（residual analysis）是确定假定的回归模型是否适宜的重要方法。

$y_i - \hat{y}_i$ 的值给出。对于 Armand 比萨饼连锁店的例子，根据表 14-7 中的数据，可以作出关于 x 的残差图。对应 $x_1 = 2$ 和 $y_1 - \hat{y}_1 = -12$，第一个点的坐标是（2，−12）；对应 $x_2 = 6$ 和 $y_2 - \hat{y}_2 = 15$，第二个点的坐标是（6，15），等等。图 14-11 是得到的残差图。

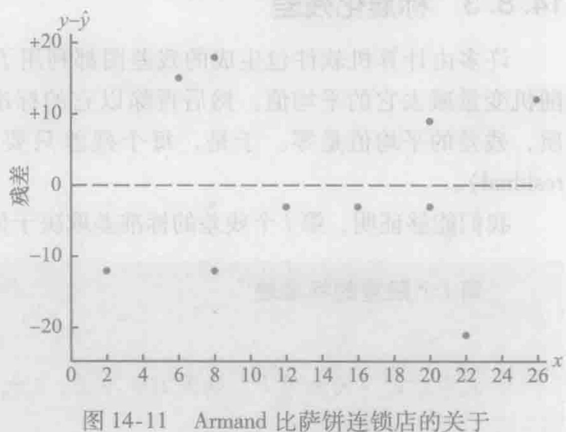


图 14-11　Armand 比萨饼连锁店的关于自变量 x 的残差图

在对这个残差图的结果进行解释之前，让我们首先考虑在任意一个残差图中可能观测到的某些一般的图形模式。在图 14-12中给出了三个例子。如果假定对所有的 x 值，ε 的方差都是相同的，并且假定描述变量 x 和 y 之间关系的回归模型是恰当的，那么残差图给出的一个总体印象是，所有的散点都应落在一条水平带中间，如图 14-12a所示。但是，如果对所有的 x 值，ε 的方差是不相同的，例如对于较大的 x 值，关于回归线的变异性也较大，如图 14-12b 所看到的那样。在这种情形下，就违背了 ε 有一个相同的常数方差的假定。另一种可能的残差图如图 14-12c所示。在这种情形下，我们的结论是，假定的回归模型不能恰当地描述变量之间的关系。这时我们应考虑曲线回归模型或者多元回归模型。

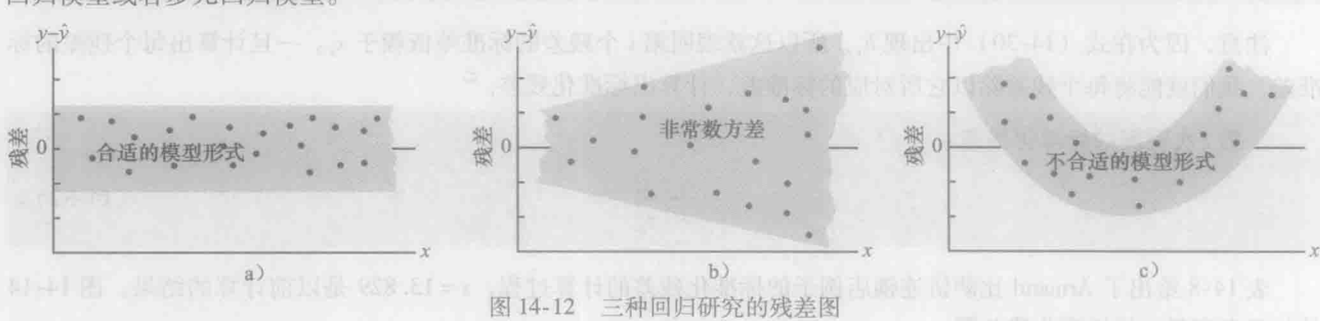


图 14-12　三种回归研究的残差图

现在让我们回到图 14-11 所示的 Armand 比萨饼连锁店的残差图。这些残差看来好像与图 14-12a 中的水平型的图形很相似。因此，我们的结论是，残差图并没有提供足够的证据，使我们对 Armand 比萨饼连锁店回归模型所作的假定表示怀疑。这时，我们有信心作出结论，Armand 比萨饼连锁店的简单线性回归模型是合理的。

在对残差图进行有效的解释时，经验和好的判断能力永远是关键因素。一个残差图与在图 14-12 中所看到的三种类型之一恰好完全一致的情形是很少见的。然而，在弄清楚合理的残差图形模式与显示模型的假定应该受到质疑的残差图形模式之间的差别时，经常从事回归研究和经常考察残差图的分析人员将成为专家。在这里介绍的残差图是用来评价回归模型假定有效性的一种方法。

14.8.2　关于$\hat{y}$的残差图

另一种残差图的做法是，用水平轴表示应变量的预测值$\hat{y}$，用纵轴表示对应的残差值。每个残差都用图上的一个点来表示。对于每一个点，第一个坐标由$\hat{y}_i$ 的值给出，第二个坐标由对应的第 i 个残差 $y_i - \hat{y}_i$ 的值给出。对于 Armand 比萨饼连锁店的例子，根据表 14-7 中的数据，对应$\hat{y}_1 = 70$ 和 $y_1 - \hat{y}_1 = -12$，第一个点的坐标是（70，−12）；对应$\hat{y}_2 = 90$ 和$y_2 - \hat{y}_2 = 15$，第二个点的坐标是（90，15），等等。图 14-13是得到的残差图。注意，这个残差图的图形模式与关于自变量 x 的残差图的图形模式是相同的。它不是让我们对模型的假定产生怀疑的那种图形模式。对于简单线性回归，关于 x 的残差图和关于$\hat{y}$的残差图，这两个图给出了同样的图形模式。对于多元回归分析，因为有一个以上的自变量，所以关于$\hat{y}$的残差图

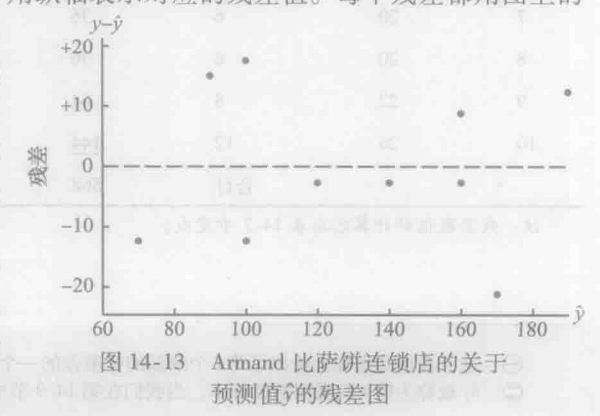


图 14-13　Armand 比萨饼连锁店的关于预测值$\hat{y}$的残差图

有着更广泛的应用。

14.8.3 标准化残差

许多由计算机软件包生成的残差图都利用了残差的标准化形式。就像我们在前面几章中看到的那样，一个随机变量减去它的平均值，然后再除以它的标准差，就得到了一个标准化的随机变量。由于最小二乘法的性质，残差的平均值是零。于是，每个残差只要简单地除以它的标准差，就得到了**标准化残差**（standardized residual）。

我们能够证明，第 i 个残差的标准差取决于估计的标准误差 s 和对应的自变量 x_i 的值。

第 i 个残差的标准差[⊖]

$$s_{y_i-\hat{y}_i} = s\sqrt{1-h_i} \tag{14-30}$$

式中，$s_{y_i-\hat{y}_i}$ 代表第 i 个残差的标准差；s 代表估计的标准误差。

$$h_i = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2} \tag{14-31}$$

注意，因为在式（14-30）中出现 h_i，所以这就表明第 i 个残差的标准差依赖于 x_i。一旦计算出每个残差的标准差，我们就能将每个残差除以它所对应的标准差，计算出标准化残差。[⊜]

第 i 次观测的标准化残差

$$\frac{y_i-\hat{y}_i}{s_{y_i-\hat{y}_i}} \tag{14-32}$$

表 14-8 给出了 Armand 比萨饼连锁店例子的标准化残差的计算过程，$s=13.829$ 是以前计算的结果。图 14-14 是关于自变量 x 的标准化残差图。

表 14-8 Armand 比萨饼连锁店的标准化残差的计算

连锁店 i	x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}$	h_i	$s_{y_i-\hat{y}_i}$	$y_i-\hat{y}_i$	标准化残差
1	2	−12	144	0.253 5	0.353 5	11.119 3	−12	−1.079 2
2	6	−8	64	0.112 7	0.212 7	12.270 9	15	1.222 4
3	8	−6	36	0.063 4	0.163 4	12.649 3	−12	−0.948 7
4	8	−6	36	0.063 4	0.163 4	12.649 3	18	1.423 0
5	12	−2	4	0.007 0	0.107 0	13.068 2	−3	−0.229 6
6	16	2	4	0.007 0	0.107 0	13.068 2	−3	−0.229 6
7	20	6	36	0.063 4	0.163 4	12.649 3	−3	−0.237 2
8	20	6	36	0.063 4	0.163 4	12.649 3	9	0.711 5
9	22	8	64	0.112 7	0.212 7	12.270 9	−21	−1.711 4
10	26	12	144	0.253 5	0.353 5	11.119 3	12	1.079 2
		合计	568					

注：残差数值的计算已在表 14-7 中完成。

⊖ 这个表达式实际上给出了第 i 个残差的标准差的一个估计，因为我们用 s 替代了 σ。

⊜ h_i 被称为第 i 次观测的杠杆率。当我们在第 14.9 节中考虑有影响的观测值时，将对杠杆率做进一步的讨论。

标准化残差图能对随机误差项 ε 服从正态分布的假定提供一种直观的认识。如果这一假定被满足，那么标准化残差的分布看起来也应该服从一个标准正态概率分布[⊖]。于是，当查看标准化残差图时，我们应该期望看到，大约有95%标准化残差介于 -2 ~ +2。从图14-14中我们看到，对于Armand比萨饼连锁店例子，所有的标准化残差都落在 -2 ~ +2。所以，根据对标准化残差图的分析，我们没有理由怀疑 ε 服从正态分布的假定。[⊜]

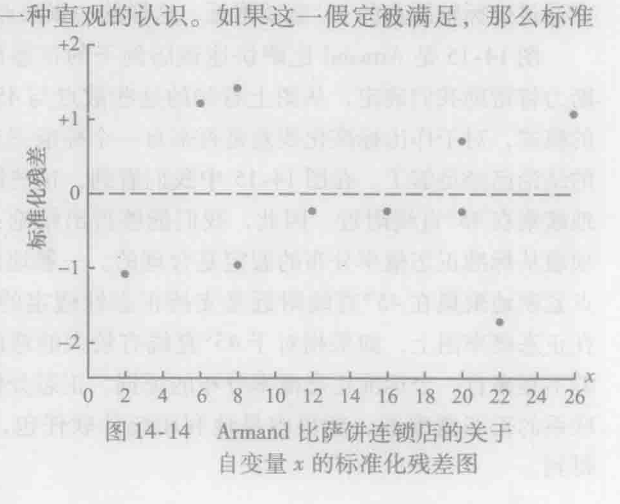


图14-14　Armand比萨饼连锁店的关于自变量 x 的标准化残差图

由于回归分析要求我们计算估计值 $\hat{y}$，残差和标准化残差，而这些计算又十分烦琐费力，所以大部分统计软件包，作为可供选择的回归分析输出，都提供了这些值的计算结果。因此，我们能很容易地得到残差图。对于较大型的回归分析问题，为了作出我们在本节讨论的残差图，计算机软件包是唯一的实用工具。

14.8.4　正态概率图

确定误差项 ε 服从正态分布的假定成立的另一个方法是**正态概率图**（normal probability plot）。为了说明如何绘制一张正态概率图，我们首先介绍正态分数的概念。

假设我们从一个平均值为0，标准差为1的标准正态概率分布中随机地抽取10个数值，并将这一抽样过程反复进行，然后把每个样本中的10个数值按从小到大的顺序排列。现在我们仅考虑每个样本中最小的数值。在反复进行的抽样过程中得到的最小值是一个随机变量，被称为一阶顺序统计量。

统计学家已经证明，来自标准正态概率分布的容量为10的样本，一阶顺序统计量的期望值为 -1.55。这个期望值被称为正态分数。对于样本容量 $n=10$ 的情形，有10个顺序统计量和10个正态分数（见表14-9）。一般地说，如果我们有一个由 n 个观测值组成的数据集，那么就有 n 个顺序统计量和 n 个正态分数。

表14-9　$n=10$ 的正态分数

顺序统计量	正态分数	顺序统计量	正态分数
1	-1.55	6	0.12
2	-1.00	7	0.37
3	-0.65	8	0.65
4	-0.37	9	1.00
5	-0.12	10	1.55

表14-10　Armand比萨饼连锁店的正态分数和排好顺序的标准化残差

正态分数	排好顺序的标准化残差	正态分数	排好顺序的标准化残差
-1.55	-1.7114	0.12	-0.2296
-1.00	-1.0792	0.37	0.7115
-0.65	-0.9487	0.65	1.0792
-0.37	-0.2372	1.00	1.2224
-0.12	-0.2296	1.55	1.4230

现在让我们说明，对于Armand比萨饼连锁店的例子，怎样利用10个正态分数去确定标准化残差服从标准正态概率分布。我们从表14-8中的10个标准化残差的排序着手。我们将10个正态分数和10个排好顺序的标准化残差放在一起，如表14-10所示。如果正态性的假定被满足，那么最小的标准化残差应接近最小的正态分数，下一个最小的标准化残差应接近下一个最小的正态分数，依此类推。如果我们用水平轴表示正态分数，用纵轴表示对应的标准化残差做一张散点图，如果标准化残差近似服从正态分布，那么在图上标出的这些散点，应密集围绕

⊖　因为我们用 s 替代式（14-30）中的 σ，所以在严格的意义上，标准化残差的概率分布，不再服从正态分布。但是在许多回归研究中，只要样本容量充分大，标准化残差的概率分布将能很好地近似服从正态分布。

⊜　稍微偏离正态性，但不会对在回归分析中应用的统计检验产生很大的影响。

在通过坐标轴原点的45°直线附近。这样的一张散点图被称为正态概率图。

图14-15是Armand比萨饼连锁店例子的正态概率图。判断力将帮助我们确定，从图上看到的这些散点与45°直线偏离的模式，对于作出标准化残差是否来自一个标准正态概率分布的结论已经足够了。在图14-15中我们看到，这些散点都紧密地聚集在45°直线附近。因此，我们能够得出结论：随机误差项服从标准正态概率分布的假定是合理的。一般地说，较多的点紧密地聚集在45°直线附近是支持正态性假定的有力证据。在正态概率图上，如果相对于45°直线有较大的弯曲，就是残差不是来自一个标准正态概率分布的证据。正态分数和与此相联系的正态概率图，能很容易地利用统计软件包，如Minitab得到。

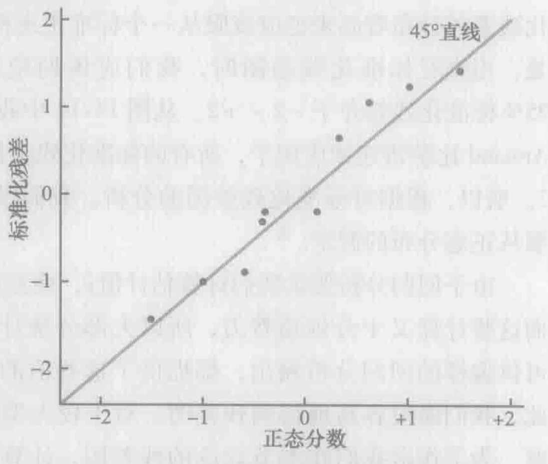


图14-15 Armand比萨饼连锁店的正态概率图

注释和评论

1. 我们利用残差和正态概率图来证实一个回归模型的假定。如果我们的检查表明一个或几个假定是不可靠的，那么我们就应该考虑一个不同的回归模型或者一个数据变换。当这些假定不成立时，一个合理适当的修正行为必须在好的判断力的基础上才能作出，来自有经验的统计学家的建议可能是宝贵的意见。
2. 统计学家用来证实回归模型的假定成立的主要方法是残差分析。即使没有发现假定被违背，但是这并不一定意味着，模型将能给出一个好的预测。然而，如果有补充的统计检验支持显著性结论，并且有比较大的判定系数，那么我们利用估计的回归方程应该能够得到好的估计值和预测值。

练 习

方法

46. 利用下面的数据进行回归研究。

观测次数	x_i	y_i	观测次数	x_i	y_i
1	2	4	6	7	6
2	3	5	7	7	9
3	4	4	8	8	5
4	5	6	9	9	11
5	7	4			

a. 对这些数据，建立估计的回归方程。

b. 绘制出残差图。关于误差项的假定你觉得被满足了吗？

应用

48. 参阅第7题，我们已经得到了一个关于售货员的工龄和年销售额的估计的回归方程。

a. 对这个问题，计算残差并绘制出残差图。

b. 根据残差图，关于随机误差项的假定你觉得合理吗？

14.9 残差分析：异常值和有影响的观测值

在第14.8节中，我们已经说明了如何利用残差分析来判定，什么时候会发生违背回归模型假定的情形。在这一节中，我们将讨论在建立估计的回归方程时，如何利用残差分析去识别异常的观测值或特别有影响的观测值。本节将详细论述，当出现这样的观测值时，我们应该采取的一些措施。

14.9.1 检测异常值

图14-16是一个只有一个**异常值**（outlier）的数据集的散点图，这个异常值是一个数据点（观测值），它与其

余的数据点所显示的趋势不相吻合。异常值表示观测值是可疑的，并且值得我们认真地检查。这些异常值可能代表了错误的数据，倘若如此，我们应该修正这些数据；这些异常值也可能意味着出现了违背模型假定的情形，倘若如此，我们应该考虑其他形式的模型；最后，这些异常值可能仅仅是由于随机因素的影响而产生的不同寻常的数值，在这种情形下，我们应该保留这些异常值。

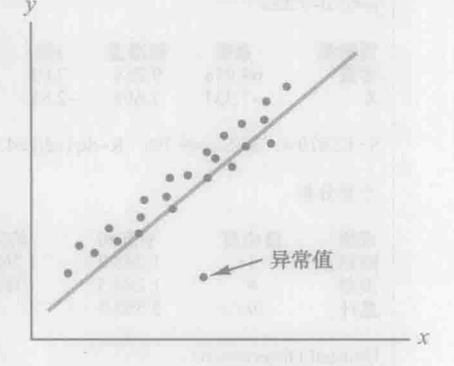
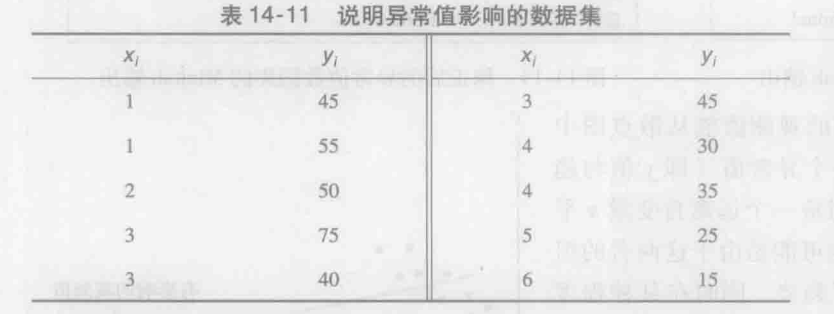


图 14-16　有一个异常值的数据集

为了说明检测异常值的过程，我们考虑表 14-11 中的数据集，图 14-17 是该数据集的散点图。除了第四个观测值（$x_4=3$，$y_4=75$）以外，散点图的形状明显地提示我们，这些数据点存在一个负线性关系。事实上，根据散点图上其余数据点的散布格局，我们可以预期 y_4 要小得多，因此我们将对应的观测值确认为是一个异常值。对于简单线性回归的情形，只要简单地检查散点图，人们往往就能探明异常值。

表 14-11　说明异常值影响的数据集



x_i	y_i	x_i	y_i
1	45	3	45
1	55	4	30
2	50	4	35
3	75	5	25
3	40	6	15

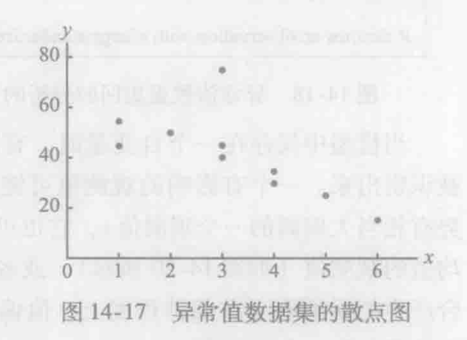


图 14-17　异常值数据集的散点图

标准化残差也能用于识别异常值。如果一个观测值与散点图上其余数据点的散布格局有一个大的偏离（例如，图 14-16 中的异常值），那么对应的标准化残差的绝对值也将是大的。许多计算机软件包能自动地识别标准化残差绝对值大的观测值。对表 14-11 的数据进行回归分析的 Minitab 输出如图 14-18 所示。图 14-18 倒数第二行的 Minitab 输出给出了第四个观测值的标准化残差是 2.67。Minitab 在输出的“Unusual Observation”（与众不同的观测值）部分，提供一个标准化残差小于 −2 或者大于 +2 的每个观测值列表。在这种情形下，这个观测值被单独打印在一行，并在标准化残差后面接着打印一个字母“R”，如图 14-18 所示。因为误差项服从正态分布，所以大约有 5% 的标准化残差应小于 −2 或者大于 +2。

在决定如何处理一个异常值时，首先我们应该核对这个异常值是不是一个有效的观测值。也许在最初记录数据时或是在将数据录入计算机文件时，已经产生了错误。例如，假定我们在核对表 14-11 中的异常值数据时，我们发现第四个观测值（$x_4=3$，$y_4=75$）出现了错误，于是将这个观测值修正为 $x_4=3$，$y_4=30$。图 14-19 是修正 y_4 的数值后得到的 Minitab 输出。我们看到，利用未经修正的数据对拟合优度有相当大的影响。而用修正后的数据，R^2 的数值从 49.7% 增加到 83.8%，b_0 的值从 64.958 减少到 59.237，回归线的斜率也从 −7.331 变成 −6.949。识别出异常值使我们能够修正数据的误差并改善回归的效果。

14.9.2　检测有影响的观测值

有时一个或几个观测值会对我们得到的回归结果产生一个强影响。图 14-20 显示的是在简单线性回归情形下存在一个**有影响的观测值**（influential observation）的例子。在图中，估计的回归线有一个负斜率。但是，如果我们将有影响的观测值从数据集中删除，那么估计的回归线的斜率将从负值改变为正值，y 轴截距将会变得比较小。显然，在确定估计的回归线时，这个观测值的影响要比任何其他观测值的影响大得多，而从数据集中删除任意一个其他的观测值，对估计的回归方程将只有很小的影响。

```
回归方程是销售额
y=65.0-7.33x

预测量    系数     标准差   t值     p-值
常数      64.958   9.258    7.02    0.000
X         -7.331   2.608    -2.81   0.023

S=12.670 4  R-Sq=49.7%  R-Sq(adj)=43.4%

方差分析

来源   自由度   平方和     均方      F      p
回归   1        1 268.2    1 268.2   7.90   0.023
误差   8        1 284.3    160.5
总计   9        2 552.5

Unusual Observations
Obs  x     y      Fit    SE Fit  Residual  St Resid
  4  3.00  75.00  42.97  4.04    32.03     2.67R

R denotes an observation with a large standardized residual.
```

图 14-18 异常值数据集回归分析的 Minitab 输出

```
回归方程是销售额
Y=59.2-6.95X

预测量    系数     标准差   t值     p-值
常数      59.237   3.835    15.45   0.000
X         -6.949   1.080    -6.43   0.000

S=5.248 08  R-Sq=83.8%  R-Sq(adj)=81.8%

方差分析

来源   自由度   平方和     均方      F       p
回归   1        1 139.7    1 139.7   41.38   0.000
误差   8        220.3      27.5
总计   9        1 360.0
```

图 14-19 修正后的异常值数据集的 Minitab 输出

当模型中仅存在一个自变量时，有影响的观测值能从散点图中被识别出来。一个有影响的观测值可能是一个异常值（即 y 值与趋势有相当大偏离的一个观测值），它也可能对应一个远离自变量 x 平均值的观测值（如图 14-20 所示），或者它也可能是由于这两者的组合产生的观测值（在某种程度上 y 值偏离了趋势，同时在某种程度上又是 x 的一个极端值）。

图 14-20 存在一个有影响的观测值的数据集

因为有影响的观测值对估计的回归方程可能有如此明显的影响，所以我们对这种观测值必须仔细检查。首先，我们应该查明在采集或者录入数据时，确保没有出现错误。如果出现了错误，我们应进行修正并求出一个新的估计的回归方程。如果一个有影响的观测值是有效的，我们应该为有这样一个观测值而感到幸运。因为这样一个有效的点，能帮助我们对模型是否合理有一个较好的理解，并且能导致我们得到一个比较好的估计的回归方程。在图 14-20 的数据集中，存在一个有影响的观测值，如果它是有效的观测值，那么将启发我们试着去得到自变量 x 的中间值的数据，从而使我们能更好地理解、认识 x 和 y 之间的关系。

自变量是极端值的观测值被称为**高杠杆率点**（high leverage point）。图 14-20 中的有影响的观测值就是一个具有高杠杆率的点。我们可以根据自变量的值与它们的平均值距离的远近确定一个观测值的杠杆率。对于只有一个自变量的情形，我们用 h_i 表示第 i 次观测的杠杆率，它能利用式（14-33）计算出来。

第 i 次观测的杠杆率

$$h_i=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2} \tag{14-33}$$

从上式中可以清晰地看出，与 x 的平均值 $\bar{x}$ 距离较远的 x_i，第 i 次观测的杠杆率较高。

作为标准回归分析输出的一部分，许多统计软件包都能自动地识别出高杠杆率的观测值。为了说明 Minitab 统计软件包如何识别具有高杠杆率的观测值，我们考虑表 14-12中的数据集。

表 14-12 具有高杠杆率观测值的数据集

x_i	y_i	x_i	y_i
10	125	20	120
10	130	25	110
15	120	70	100
20	115		

图 14-21 是表 14-12 数据集的散点图。从图中可以清晰地看出，第 7 个观测值（$x=70$，$y=100$）是自变量 x 一

个具有极端值的观测值。因此，我们将期待着它作为高杠杆率的点被识别出来。对于这个观测值，我们利用式（14-33）计算出它的杠杆率如下。

$$h_7 = \frac{1}{n} + \frac{(x_7 - \bar{x})^2}{\sum(x_i - \bar{x})^2} = \frac{1}{7} + \frac{(70 - 24.286)^2}{2\,621.43} = 0.94$$

对于简单线性回归情形，如果杠杆率 $h_i > \min\{6/n, 0.99\}$（即 h_i 大于 $6/n$ 与 0.99 的较小者），那么 Minitab 将把这个观测值识别为具有高杠杆率的观测值。对于表 14-12 中的数据集，$6/n = 6/7 = 0.86$，因为 $h_7 = 0.94 > 0.86$，所以，Minitab 将把第 7 个观测值识别为自变量 x 的值具有大的影响的观测值。图 14-22 给出了这个数据集回归分析的 Minitab 输出。第 7 个观测值（$x = 70$，$y = 100$）作为具有大的影响的观测值被识别出来，这个具有大的影响的观测值在图的底部被单独打印一行，并在这行的右侧边缘处用“X”标出。[⊖]

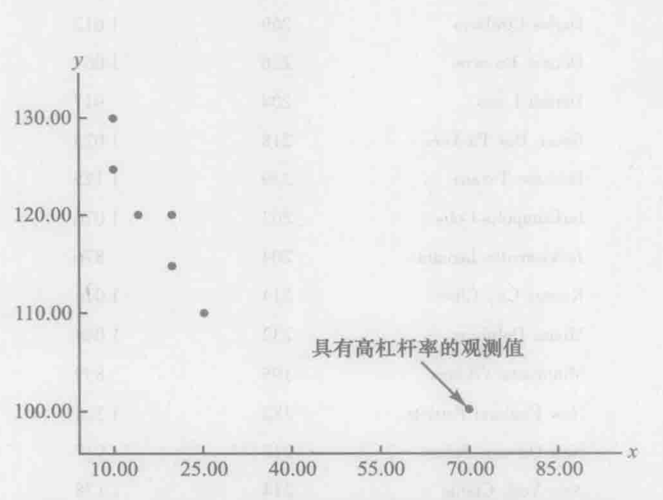


图 14-21　具有一个高杠杆率观测值的数据集的散点图

回归方程是销售额
y=127−0.425x

预测量	系数	标准差	t值	p-值
常数	127.466	2.961	43.04	0.000
X	−0.425 07	0.095 37	−4.46	0.007

S=4.882 82　R-sq=79.9%　R-sq (adj) =75.9%

方差分析

来源	自由度	平方和	均方	F	p
回归	1	473.65	473.65	19.87	0.007
误差	5	119.21	23.84		
总计	6	592.86			

Unusual Observations

Obs	x	y	Fit	SE Fit	Residual	St Resid
7	70.0	100.00	97.71	4.73	2.29	1.91 X

X denotes an observation whose X value gives it large influence.

图 14-22　具有一个高杠杆率观测值的数据集的 Minitab 输出

有影响的观测值是由于大的残差和高杠杆率的交互作用而产生的，我们要把它们检测出来可能是一件困难的工作。在确定什么时候一个观测值是有影响的观测值时，只要考虑到大的残差和高杠杆率这两个方面，我们就能得到便于使用的判断方法。其中一种这样的度量称为库克 D 统计量，我们将在第 15 章中加以讨论。

注释和评论

一旦一个观测值由于大的残差或高杠杆率被识别为是潜在的有影响的观测值，我们就应该针对这个观测值对估计的回归方程的影响进行评估。较高级的教科书将对此进行讨论。然而，如果人们不熟悉较高级教科书的内容，那么一个简单的，不管包含或是不包含有影响观测值都适用的步骤将用于回归分析。这个方法将显示出观测值对分析结果的影响。

练　习

方法

☆50. 考虑下面两个变量 x 和 y 的数据。

x_i	135	110	130	145	175	160	120
y_i	145	100	120	120	130	130	110

a. 计算这些数据的标准化残差。在这些数据中包含任何异常值吗？请作出解释。

b. 作出关于 $\hat{y}$ 的标准化残差图。这个残差图显示出任何异常值吗？

c. 绘制出这些数据的散点图。散点图表明在这些数据中存在任何异常值吗？在一般情况下，对

⊖ 为了完成识别有影响的观测值的计算，计算机软件包是必不可少的。在这里我们讨论了 Minitab 的选择规则。

于简单线性回归，这一发现意味着什么？

应用

☆52. Charity Navigator 是美国领先的独立慈善评估机构。对于10个超大型慈善机构的总费用（单位：美元）、行政费用占总预算支出的比例、筹款费用占总预算支出的比例，以及开展活动费用占总预算支出的比例的数据如下表所示（Charity Navigator website，2012年4月12日）。行政费用包括管理费、行政人员及相关费用以及组织会议的费用。筹款费用是慈善机构花在募捐上的费用，开展活动费用是慈善机构花在策划活动方案和提供服务上的费用。由于四舍五入的原因，这三个百分比之和不等于100%。

慈善机构	总费用（美元）	行政费用（%）	筹款费用（%）	活动费用（%）
American Red Cross	3 354 177 445	3.9	3.8	92.1
World Vision	1 205 887 020	4.0	7.5	88.3
Smithsonian Institution	1 080 995 083	23.5	2.6	73.7
Food For The Poor	1 050 829 851	0.7	2.4	96.8
American Cancer Society	1 003 781 897	6.1	22.2	71.6
Volunteers of America	929 158 968	8.6	1.9	89.4
Dana-Farber Cancer Institute	877 321 613	13.1	1.6	85.2
AmeriCares	854 604 824	0.4	0.7	98.9
ALSAC-St. Jude Children's Research Hospital	829 662 076	9.6	16.9	73.4
City of Hope	736 176 619	13.7	3.0	83.1

a. 用横轴表示筹款费用所占比例（%），用纵轴表示活动费用所占比例（%），绘制出这些数据的散点图。观察这些数据，是否出现任何的异常值和（或）有影响的观测值？

b. 建立这些数据的估计的回归方程，使这个方程在筹款费用所占比例（%）已知时，能用来预测开展活动费用所占比例（%）。

c. 对于这个问题，估计的回归方程斜率的数值是否有意义？

d. 利用残差分析来确定：是否存在任何的异常值和（或）有影响的观测值？简要地概述一下你的发现和结论。

54. 下面是美国橄榄球联盟（NFL）32支球队的年收入（100万美元）和估计的球队商业价值（100万美元）的统计资料（Forbes website，2009年2月）。

球队名称	年收入（100万美元）	球队价值（100万美元）
Arizona Cardinals	203	914
Atlanta Falcons	203	872
Baltimore Ravens	226	1 062
Buffalo Bills	206	885
Carolina Panthers	221	1 040
Chicago Bears	226	1 064
Cincinnati Bengals	205	941
Cleveland Browns	220	1 035
Dallas Cowboys	269	1 612
Denver Broncos	226	1 061
Detroit Lions	204	917
Green Bay Packers	218	1 023
Houston Texans	239	1 125
Indianapolis Colts	203	1 076
Jacksonville Jaguars	204	876
Kansas City Chiefs	214	1 016
Miami Dolphins	232	1 044
Minnesota Vikings	195	839
New England Patriots	282	1 324
New Orleans Saints	213	937
New York Giants	214	1 178
New York Jets	213	1 170
Oakland Raiders	205	861
Philadelphia Eagles	237	1 116
Pittsburgh Steelers	216	1 015
San Diego Chargers	207	888
San Francisco 49ers	201	865
Seattle Seahawks	215	1 010
St. Louis Rams	206	929
Tampa Bay Buccaneers	224	1 053
Tennessee Titans	216	994
Washington Redskins	327	1 538

a. 用水平轴表示球队的年收入，用纵轴表示球队的商业价值，绘制出这些数据的散点图。查看这张散点图，在这些数据里，存在任何的异常值和（或）有影响的观测值吗？

b. 建立估计的回归方程，使这个方程在球队的年收入已知时，能用来预测球队的商业价值。

c. 应用残差分析方法确定：是否存在任何异常值和（或）有影响的观测值。简要地总结一下你的发现和结论。

小结

在这一章里，我们说明了回归分析如何用于确定一个应变量 y 怎样依赖一个自变量 x。在简单线性回归情形，回归模型是 $y=\beta_0+\beta_1x+\varepsilon$。简单线性回归方程 $E(y)=\beta_0+\beta_1x$ 描述了 y 的平均值或期望值是如何依赖 x 的。我们利用样本数据和最小二乘法建立了估计的回归方程 $\hat{y}=b_0+b_1x$。实际上，b_0 和 b_1 是用来估计模型的未知参数 β_0 和 β_1 的样本统计量。

作为估计的回归方程拟合优度的一个度量，我们介绍了判定系数的概念。判定系数是应变量 y 中的变异性能被估计的回归方程解释的部分所占的比例。我们还回顾了相关系数的概念，它是描述两变量之间线性相关强度的一个度量。

我们讨论了回归模型及与其相联系的随机误差项 ε 的假定，并且介绍了建立在这些假定基础上的 t 检验和 F 检验，这两个检验都是判断两变量之间的关系在统计上是否显著的工具。我们说明了如何利用估计的回归方程建立 y 的平均值的置信区间估计和 y 的个别值的预测区间估计的方法。

作为本章的结束，我们用一节篇幅介绍了回归问题的计算机解法，用两节篇幅介绍了利用残差分析验证模型的假定和识别异常值及有影响的观测值的方法。

关键术语

dependent variable 应变量 是被预测或被解释的变量，用 y 表示。

independent variable 自变量 是用于预测或解释的变量，用 x 表示。

simple linear regression 简单线性回归 含有一个自变量和一个应变量，并且两个变量之间的关系用一条直线近似的回归分析。

regression model 回归模型 描述应变量 y 如何依赖自变量 x 和随机误差项的方程。对于简单线性回归情形，回归模型是 $y=\beta_0+\beta_1x+\varepsilon$。

regression equation 回归方程 描述应变量 y 的平均值或期望值如何依赖自变量 x 的方程。对于简单线性回归情形，回归方程是 $E(y)=\beta_0+\beta_1x$。

estimated regression equation 估计的回归方程 利用最小二乘法，根据样本数据建立的回归方程的估计。对于简单线性回归情形，估计的回归方程是 $\hat{y}=b_0+b_1x$。

least squares method 最小二乘法 用建立估计的回归方程的方法，其目标是使式 $\sum(y_i-\hat{y}_i)^2$ 极小化。

scatter diagram 散点图 用水平轴表示自变量，用纵轴表示应变量的二维数据图。

Coefficient of determination 判定系数 估计的回归方程拟合优度的度量。该系数能被理解为应变量 y 的变异性能被估计的回归方程解释的部分所占的比例。

***i* th residual 第 *i* 个残差** 应变量的观测值与利用估计的回归方程得到的预测值之间的差；对于第 i 个观测值，第 i 个残差是 $y_i-\hat{y}_i$。

correlation coefficient 相关系数 两变量之间线性关系强度的度量（在前面第3章中已讨论过）。

mean square error 均方误差 随机误差项 ε 的方差 σ^2 的无偏估计量，用 MSE 或 s^2 表示。

standard error of the estimate 估计的标准误差 均方误差的平方根，用 s 表示。它是随机误差项 ε 的标准差 σ 的估计。

ANOVA table ANOVA表 方差分析表，用于概括与显著性 F 检验相联系的计算。

confidence interval 置信区间 对于一个给定的 x 值，y 的平均值的区间估计。

prediction interval 预测区间 对于一个给定的 x 值，y 的一个个别值的区间估计。

residual analysis 残差分析 残差分析是用于确定关于回归模型所做的假定是否成立的分析方法。残差分析还被用于识别异常值和有影响的观测值。

residual plot 残差图 残差的图形表示，用于确定关于回归模型所做的假定是否成立的一种图示方法。

standardized residual 标准化残差 残差除以它的标准差后得到的数值。

normal probability plot 正态概率图 标准化残差关于正态分数值绘制出的散点图。正态概率图能帮助我们确定随机误差项服从正态概率分布的假定是否成立。

outlier 异常值 与其余的数据点所显示的趋势不相吻合的数据点或观测值。

influential observation 有影响的观测值 对回归结果有一个强影响或强作用的观测值。

high leverage points 高杠杆率点 自变量的值是极端值的观测值。

重要公式

简单线性回归模型 $y=\beta_0+\beta_1 x+\varepsilon$ (14-1)

简单线性回归方程 $E(y)=\beta_0+\beta_1 x$ (14-2)

估计的简单线性回归方程 $\hat{y}=b_0+b_1 x$ (14-3)

最小二乘准则 $\min\sum(y_i-\hat{y}_i)^2$ (14-5)

估计的回归方程的斜率和 y 轴截距

$$b_1=\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} \quad (14\text{-}6)$$

$$b_0=\bar{y}-b_1\bar{x} \quad (14\text{-}7)$$

误差平方和 $SSE=\sum(y_i-\hat{y}_i)^2$ (14-8)

总平方和 $SST=\sum(y_i-\bar{y})^2$ (14-9)

回归平方和 $SSR=\sum(\hat{y}_i-\bar{y})^2$ (14-10)

SST，SSR 和 SSE 之间的关系

$$SST = SSR + SSE \quad (14\text{-}11)$$

判定系数 $r^2=\frac{SSR}{SST}$ (14-12)

样本相关系数 $r_{xy}=(b_1\text{ 的符号})\sqrt{\text{判定系数}}$

$$=(b_1\text{ 的符号})\sqrt{r^2} \quad (14\text{-}13)$$

均方误差（σ^2 的估计量）

$$s^2=MSE=\frac{SSE}{n-2} \quad (14\text{-}15)$$

估计的标准误差 $s=\sqrt{MSE}=\sqrt{\frac{SSE}{n-2}}$ (14-16)

b_1 的标准差 $\sigma_{b_1}=\frac{\sigma}{\sqrt{\sum(x_i-\bar{x})^2}}$ (14-17)

b_1 的估计的标准差 $s_{b_1}=\frac{s}{\sqrt{\sum(x_i-\bar{x})^2}}$ (14-18)

t 检验统计量 $t=\frac{b_1}{s_{b_1}}$ (14-19)

均方回归 $MSR=\frac{SSR}{\text{自变量的个数}}$ (14-20)

F 检验统计量 $F=\frac{MSR}{MSE}$ (14-21)

$\hat{y}^*$ 的估计的标准差

$$s_{\hat{y}^*}=s\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \quad (14\text{-}23)$$

$E(y^*)$ 的置信区间 $\hat{y}^*\pm t_{\alpha/2}s_{\hat{y}^*}$ (14-24)

一个个别值的估计的标准差

$$s_{\text{pred}}=s\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \quad (14\text{-}26)$$

y^* 的预测区间 $\hat{y}^*\pm t_{\alpha/2}s_{\text{pred}}$ (14-27)

第 i 个观测值的残差 $y_i-\hat{y}_i$ (14-28)

第 i 个残差的标准差 $s_{y_i-\hat{y}_i}=s\sqrt{1-h_i}$ (14-30)

第 i 个观测值的标准化残差 $\frac{y_i-\hat{y}_i}{s_{y_i-\hat{y}_i}}$ (14-32)

第 i 个观测值的杠杆率

$$h_i=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2} \quad (14\text{-}33)$$

补充练习

56. 对于 x 的一个给定值，y 的平均值的区间估计和对于 x 的一个给定值，y 的一个个别值的区间估计，请用你自己的话对它们两者之间的区别作出解释。

58. 道琼斯工业平均指数（DJIA）和标准普尔 500 指数（S&P 500）是用于度量股市整体变动趋势的测度。道琼斯工业平均指数是以 30 家大型企业的价格变动为依据；标准普尔 500 指数是由 500 只股票组成的一个综合指数。有人说，标准普尔 500 指数是表现股市业绩的一个更好的测度，因为它的基础更广泛。从 2012 年 1 月 6 日开始，15 周的道琼斯工业平均指数和标准普尔 500 指数的收盘价如下表所示（*Barron's* website，2012 年 4 月 17 日）。

日期	DJIA	S&P 500
1 月 6 日	12 360	1 278
1 月 13 日	12 422	1 289
1 月 20 日	12 720	1 315
1 月 27 日	12 660	1 316
2 月 3 日	12 862	1 345
2 月 10 日	12 801	1 343
2 月 17 日	12 950	1 362
2 月 24 日	12 983	1 366
3 月 2 日	12 978	1 370
3 月 9 日	12 922	1 371
3 月 16 日	13 233	1 404
3 月 23 日	13 081	1 397
3 月 30 日	13 212	1 408
4 月 5 日	13 060	1 398
4 月 13 日	12 850	1 370

a. 以DJIA为自变量，绘制出这些数据的散点图。

b. 建立一个估计的回归方程。

c. 在$\alpha=0.05$的显著性水平下，对两个变量之间的关系进行显著性检验。

d. 估计的回归方程对观测数据的拟合好吗？请作出解释。

e. 假定DJIA的收盘价是13 500。预测S&P 500的收盘价。

f. 我们是否应该关注，在（e）中用来预测S&P 500收盘价的DJIA收盘价13 500，因为该数值13 500已经超出了在建立估计的回归方程时使用的数据范围。

60. 近年来，高等教育最大的一个变化是远程网络大学教育的发展。在线教育数据库（Online Education Database）是一个独立机构，该机构的宗旨是制定一个经过认证的顶级的远程网络学院的详细清单。下表给出了29所远程网络学院的保留率（%）和毕业率（%）的统计数据（Online Education Database website，2009年1月）

远程网络学院名称	保留率(%)	毕业率(%)
Western International University	7	25
South University	51	25
University of Phoenix	4	28
American InterContinental University	29	32
Franklin University	33	33
Devry University	47	33
Tiffin University	63	34
Post University	45	36
Peirce College	60	36
Everest University	62	36
Upper Iowa University	67	36
Dickinson State University	65	37
Western Governors University	78	37
Kaplan University	75	38
Salem International University	54	39
Ashford University	45	41
ITT Technical Institute	38	44
Berkeley College	51	45
Grand Canyon University	69	46
Nova Southeastern University	60	47
Westwood College	37	48
Everglades University	63	50
Liberty University	73	51
LeTourneau University	78	52
Rasmussen College	48	53
Keiser University	95	55
Herzing College	68	56
National University	100	57
Florida National College	100	61

a. 以保留率为自变量，绘制出这些数据的散点图。散点图显示出在这两个变量之间存在什么关系？

b. 建立估计的回归方程。

c. 在$\alpha=0.05$的显著性水平下，检验变量之间关系的显著性。

d. 估计的回归方程对观测数据的拟合好吗？

e. 假设你是South University的校长。在详细审查上表中的排行榜并且与其他远程网络学院比较后，你是否关注你所在学校的业绩？

f. 假设你是University of Phoenix的校长。在详细审查上表中的排行榜并且与其他网络远程学院比较后，你是否关注你所在学校的业绩？

62. 在生产过程中，管理人员认为：在检验过程中，装配线的速度（英尺/分）会影响发现次品数量的多少。为了验证这种说法，管理人员设计了一种装置，该装置能使同一批次产品，在装配线的各种速度下接受检验。收集到的数据如下表所示。

装配线的速度（英尺/分）	被发现的次品数量	装配线的速度（英尺/分）	被发现的次品数量
20	21	30	16
20	19	60	14
40	15	40	17

a. 建立被发现的次品数量关于装配线速度的估计的回归方程。

b. 在$\alpha=0.05$的显著性水平下，确定装配线的速度和被发现的次品数量之间是否相关？

c. 估计的回归方程对观测数据的拟合好吗？

d. 如果一条装配线的速度为50英尺/分，建立一个能预测次品平均数量的置信水平为95%的置信区间。

64. 对于一个大都市城区，当地交通管理部门想要确定公共汽车的使用时间和年维修费用之间是否存在某种关系。由10辆公共汽车组成一个样本，收集到的数据如下表所示。

公共汽车使用时间（年）	年维修费用（美元）	公共汽车使用时间（年）	年维修费用（美元）
1	350	3	550
2	370	4	750
2	480	4	800
2	520	5	790
2	590	5	950

a. 利用最小二乘法，建立估计的回归方程。

b. 在 $\alpha=0.05$ 的显著性水平下，通过检验能否看出两变量之间存在一个显著的关系。

c. 最小二乘回归线对观测数据的拟合好吗？请作出解释。

d. 如果有一辆已使用了4年的特定的公共汽车，建立一个该公共汽车置信水平为95%的年维修费用的预测区间。

66. 据路透社报道，施乐公司的市场贝塔系数是1.22（Reuters website，2009年1月30日）。对于个别股票，贝塔系数是由简单线性回归模型确定的。对于每一只股票，应变量是该股票的季度回报率（资本增值加上分红）减去可以从无风险投资得到的回报率（用美国国库债券率作为无风险利率）。自变量是股票市场（S&P 500）的季度回报率（资本增值加上分红）减去无风险投资回报率。利用季度数据建立估计的回归方程；估计的回归方程的斜率（b_1）就是股票市场的贝塔系数。经常将市场贝塔系数的值视为对股票风险的度量。如果市场贝塔系数的值大于1，表示这只股票的波动较市场的平均水平大；如果市场贝塔系数的值小于1，表示这只股票的波动较市场的平均水平小。假设，下面是S&P 500和Horizon Technology公司的10个季度的回报率与无风险回报率之差的统计数据。

S&P 500	Horizon	S&P 500	Horizon
1.2	−0.7	1.2	4.1
−2.5	−2.0	3.0	2.6
−3.0	−5.5	−1.0	2.0
2.0	4.7	0.5	−1.3
5.0	1.8	2.5	5.5

a. 建立估计的回归方程，使这个方程能用来预测Horizon Technology公司的市场贝塔系数。Horizon Technology公司的市场贝塔系数是多少？

b. 在 $\alpha=0.05$ 的显著性水平下，检验这两个变量之间的显著性关系。

c. 估计的回归方程对观测数据的拟合好吗？请作出解释。

d. 根据施乐公司和Horizon Technology公司的市场贝塔系数值，比较这两只股票的风险。

68. 丰田凯美瑞是北美最畅销的车型之一。二手凯美瑞汽车的销售价格取决于许多因素，包括车型年份、行驶里程和车况。为了调查汽车行驶里程和销售价格之间的关系，由19辆已销售的2007年款凯美瑞汽车组成一个样本，它们的行驶里程和销售价格数据如下表所示（PriceHub，2012年2月24日）。

a. 以汽车行驶里程为水平轴，汽车销售价格为纵轴，绘制出这些数据的散点图。

b. 根据在（a）中作出的散点图，在这两个变量之间显示出什么关系？

c. 建立估计的回归方程，使这个方程在汽车行驶里程（单位：1 000英里）已知时，能用来预测汽车销售价格（单位：1 000美元）。

d. 在 $\alpha=0.05$ 的显著性水平下，检验这两个变量之间的显著关系。

e. 估计的回归方程对观测数据的拟合好吗？请作出解释。

f. 请对估计的回归方程的斜率作出解释。

g. 假定，你正在考虑购买一辆2007年款的凯美瑞二手车，该车已经行驶了60 000英里。利用在（c）中建立的估计的回归方程，预测这辆车的价格。这是你打算支付给经销商的价格吗？

行驶里程（1 000英里）	价格（1 000美元）	行驶里程（1 000英里）	价格（1 000美元）
22	16.2	110	8.3
29	16.0	28	12.5
36	13.8	59	11.1
47	11.5	68	15.0
63	12.5	68	12.2
77	12.9	91	13.0
73	11.2	42	15.6
87	13.0	65	12.7
92	11.8	110	8.3
101	10.8		

案例14-1 测量股票市场风险

一只个别股票的风险或者波动性的的测度是过去几个时期总回报率（资本增值加上分红）的标准差。虽然计算标准差是一件容易的事情，但是作为标准市场指数（例如S&P 500）的一个函数，并没有考虑到某一股票价格变化的范围。结果是，许多金融分析师更喜欢使用另一种被称为贝塔系数的风险测度。

一只个别股票的贝塔系数是由简单线性回归模型确定的。应变量是该股票的总回报率，自变量是股票市场的总回报率。[⊖]对于这种情况下的问题，我们将利用S&P 500指数作为股票市场的总回报率的测度，并且利用月度数据，建立一个估计的回归方程。估计的回归方程的斜率（b_1）就是股票的贝塔系数。在本书所附光盘中名为Beta的文件提供了交易量大的8只普通股票和S&P 500的36个月度的总回报率（资本增值加上分红）的数据。

股票市场的贝塔值始终为1；因此，随着股票市场的上升和下降，股票也将有一个接近于1的贝塔值。如果贝塔值大于1，表示这只股票的波动较股票市场的平均水平大；如果贝塔值小于1，表示这只股票的波动较股票市场的平均水平小。例如，如果一只股票的贝塔值是1.4，这就意味着，这只股票的波动较股票市场的平均水平高40%；如果一只股票的贝塔值是0.4，这就意味着，这只股票的波动较股票市场的平均水平低60%。

管理报告

你被指派来对这些股票进行风险特性分析。编写一份报告，该报告包括但不限于下面的项目。

1. 对每一只股票和S&P 500，计算描述统计量。评论你的结果。哪一只股票的波动性最大？
2. 计算每一只股票的贝塔值。在一个上升的股票市场上，你预期哪一只股票将会有最好的业绩？在一个下跌的股票市场上，你预期哪一只股票的保值将会最佳？
3. 一只个别股票的回报率有多少能被股票市场解释。

案例14-2 美国交通部

作为交通安全研究的一部分，美国交通部在由42个城市组成的样本中，收集了每1 000个有驾驶执照的司机发生死亡事故的车祸次数和有驾驶执照的司机中21岁以下者所占比例的数据。在一年期间收集的数据如下表所示。这些数据存放在本书所附光盘名为Safety的文件中。

司机中21岁以下者所占比例（%）	每千个有驾照司机中发生死亡事故的次数	司机中21岁以下者所占比例（%）	每千个有驾照司机中发生死亡事故的次数	司机中21岁以下者所占比例（%）	每千个有驾照司机中发生死亡事故的次数	司机中21岁以下者所占比例（%）	每千个有驾照司机中发生死亡事故的次数
13	2.962	16	2.801	8	2.190	18	3.614
12	0.708	12	1.405	16	3.623	10	1.926
8	0.885	9	1.433	15	2.623	14	1.643
12	1.652	10	0.039	9	0.835	16	2.943
11	2.091	9	0.338	8	0.820	12	1.913
17	2.627	11	1.849	14	2.890	15	2.814
18	3.830	12	2.246	8	1.267	13	2.634
8	0.368	14	2.855	15	3.224	9	0.926
13	1.142	14	2.352	10	1.014	17	3.256
8	0.645	11	1.294	10	0.493		
9	1.028	17	4.100	14	1.443		

[⊖] 为了计算贝塔系数，各种模型使用了不同的方法。例如，一些模型在计算估计的回归方程以前，要从应变量和自变量中减去可以从无风险投资（例如，国库券）得到的回报率。有些模型还利用了股票市场总回报率的不同的指数；例如，利用纽约证券交易所综合指数的价值线（Value Line）计算贝塔系数。

管理报告

1. 对这些数据作出数值的和图形的概述。

2. 利用回归分析方法研究发生死亡事故的车祸次数和司机中21岁以下者所占比例之间的关系。对你的研究结果进行讨论。

3. 从你的分析中，你能得到什么结论或提出什么建议吗？

案例14-3 挑选一台傻瓜型（point-and-shoot）数码相机

《消费者报告》杂志测试了166台不同的傻瓜数码相机。基于如：分辨率（单位：100万像素）、重量（单位：盎司）、图像质量和易用性等诸多因素，该杂志对每台被测试的相机给出了一个总体得分。总分数范围为0~100，分数越高表示整体测试结果越好。因为有许多选项，选择一台相机可能是一个艰难的过程，并且对于大多数消费者来说，价格肯定是一个关键问题。通过花更多的钱，消费者将确实得到一台优越的相机吗？像素常常被认为是衡量画面质量好坏的一个重要因素，那么像素较多的相机比像素较少的相机更贵吗？由《消费者报告》杂志测试的13台佳能和15台尼康超薄型数码相机的品牌、平均零售价（美元）、分辨率（100万像素）、重量（盎司）以及总体得分数据如表14-13所示（*Consumer Reports* website，2012年2月7日）

表14-13 28台傻瓜型（point-and-shoot）数码相机的数据

观测序号	品牌	价格（美元）	分辨率（100万像素）	重量（盎司）	分数	观测序号	品牌	价格（美元）	分辨率（100万像素）	重量（盎司）	分数
1	佳能	330	10	7	66	15	尼康	300	16	7	63
2	佳能	200	12	5	66	16	尼康	200	14	6	61
3	佳能	300	12	7	65	17	尼康	400	14	7	59
4	佳能	200	10	6	62	18	尼康	120	14	5	57
5	佳能	180	12	5	62	19	尼康	170	16	6	56
6	佳能	200	12	7	61	20	尼康	150	12	5	56
7	佳能	200	14	5	60	21	尼康	230	14	6	55
8	佳能	130	10	7	60	22	尼康	180	12	6	53
9	佳能	130	12	5	59	23	尼康	130	12	6	53
10	佳能	110	16	5	55	24	尼康	80	12	7	52
11	佳能	90	14	5	52	25	尼康	80	14	7	50
12	佳能	100	10	6	51	26	尼康	100	12	4	46
13	佳能	90	12	7	46	27	尼康	110	12	5	45
14	尼康	270	16	5	65	28	尼康	130	14	4	42

管理报告

1. 对这些数据作出数值概述。

2. 以总体得分为应变量，绘制三个散点图，一个以价格为自变量，一个以分辨率为自变量，一个以重量为自变量。这三个自变量中的哪一个似乎是总体得分的最佳预测变量？

3. 利用简单线性回归分析方法，建立一个估计的回归方程，使这个方程在相机的价格已知时，能用来预测相机的总体得分。对于这个估计的回归方程，进行残差分析，并且讨论你的发现和结论。

4. 仅仅使用佳能相机的观测值进行数据分析。讨论使用简单线性回归的适用性，并且提出仅仅使用相机价格作出有关预测总体得分的任何建议。

案例14-4 找到最合适的汽车价值

当你试图决定购买一辆汽车时，决定汽车实际价值的并不一定是你在初次购买时所花的钱数。相反，性能可靠的并且不用花很多钱就能买到的汽车，往往代表了最合适的价值。但是，不管汽车的性能多么可靠，或者价钱多么便宜就能买到，它还必须要有很好的操控能力。

为了度量汽车的价值，《消费者报告》杂志提出了一个被称为价值分数的统计量。价值分数是根据汽车车主五年的费用、汽车道路综合测试分数以及预测可靠性等级得出的。车主五年的费用是该车在第一个五年所产生的各项支出，包括折旧、燃油、维修及保养，等等。根据一辆车每年行驶12 000英里的全国平均水平，得到每英里的平均费用作为车主五年费用的测度。道路测试分数是超过50次测试和评估的结果，并且以100分为满分来计算成绩，较高的分数表示有较好的性能、舒适性、便利性及燃油经济性。由《消费者报告》杂志所实施的道路测试，得到最高分的是一辆雷克萨斯LS 460L，为99分。预测可靠性等级（1=差，2=一般，3=好，4=很好，5=优秀）是基于*Consumer Reports*' Annual Auto Survey的数据得到的。

价值分数为1.0的汽车被认为是“平均值”。价值分数为2.0的汽车被认为是比价值分数为1.0的汽车好两倍的汽车；价值分数为0.5的汽车被认为只有价值分数为1.0的汽车一半好的汽车；等等。20辆被测试的家庭轿车数据，包括每一辆车的价格（美元）如下表所示。

汽车品牌和型号	价格（美元）	费用/英里	道路测试分	预测可靠性等级	价值分数
Nissan Altima 2.5S(4-cyl.)	23 970	0.59	91	4	1.75
Kia Optima LX(2.4)	21 885	0.58	81	4	1.73
Subaru Legacy 2.5i Premium	23 830	0.59	83	4	1.73
Ford Fusion Hybrid	32 360	0.63	84	5	1.70
Honda Accord LX-P(4-cyl.)	23 730	0.56	80	4	1.62
Mazda6 i Sport(4-cyl.)	22 035	0.58	73	4	1.60
Hyundai Sonata GLS(2.4)	21 800	0.56	89	3	1.58
Ford Fusion SE(4-cyl.)	23 625	0.57	76	4	1.55
Chevrolet Malibu LT(4-cyl.)	24 115	0.57	74	3	1.48
Kia Optima SX(2.0T)	29 050	0.72	84	4	1.43
Ford Fusion SEL(V6)	28 400	0.67	80	4	1.42
Nissan Altima 3.5 SR(V6)	30 335	0.69	93	4	1.42
Hyundai Sonata Limited(2.0T)	28 090	0.66	89	3	1.39
Honda Accord EX-L(V6)	28 695	0.67	90	3	1.36
Mazda6 s Grand Touring(V6)	30 790	0.74	81	4	1.34
Ford Fusion SEL(V6, AWD)	30 055	0.71	75	4	1.32
Subaru Legacy 3.6R Limited	30 094	0.71	88	3	1.29
Chevrolet Malibu LTZ(V6)	28 045	0.67	83	3	1.20
Chrysler 200 Limited(V6)	27 825	0.70	52	5	1.20
Chevrolet Impala LT(3.6)	28 995	0.67	63	3	1.05

管理报告

1. 对这些数据作出数值概述。

2. 利用回归分析方法建立一个估计的回归方程，使这个方程在汽车的价格已知时，能用来预测该车的价值分数。

3. 利用回归分析方法建立一个估计的回归方程，使这个方程在汽车车主五年成本（费用/英里）已知时，能用来预测该车的价值分数。

4. 利用回归分析方法建立一个估计的回归方程，使这个方程在汽车的道路测试分数已知时，能用来预测该车的价值分数。

5. 利用回归分析方法建立一个估计的回归方程，使这个方程在汽车的预测可靠性等级已知时，能用来预测该车的价值分数。

6. 从你的分析中，你能得出什么结论？

第15章 多元回归

CHAPTER 15

实践中的统计

dunnhumby ⊖

英国，伦敦

1989年，Clive Humby（数学家）和Edwina Dunn（营销人员）夫妇的团队创立了dunnhumby公司。dunnhumby公司将经过证实的与生俱来的能力与宏大的目标结合起来，以寻找客户购买什么样的商品以及为什么购买这些商品的线索和模式。该公司将这些见解变成可操作的策略，这些策略创造了急剧增长和可持续的忠诚度，最终提高了品牌价值和客户的认知度。

在欧洲、亚洲和美洲，dunnhumby雇用的员工超过950人，该公司为一些声望很高的著名公司提供服务，如克罗格、乐购、可口可乐，通用磨坊、金佰利、百事、宝洁和家得宝。dunnhumby的美国分公司是克罗格公司与dunnhumby组成的一家合资企业，在纽约、芝加哥、亚特兰大、明尼阿波利斯、辛辛那提和波特兰设有办事处。

公司的研究工作始于收集消费者的统计数据。这些数据来自于客户的奖励或优惠卡购买记录，以及电子销售点的交易和传统的市场研究。数据的分析常常要将数以十亿计的数据点转换成有关消费者的行为、喜好和生活方式的详细见解。这种见解有利于制订更有效的营销方案，包括定价、促销、广告、产品分类决策的战略建议。

研究人员使用被称为logistic回归的多元回归技术帮助他们分析以客户为基础的数据。使用logistic回归，建立一个估计的多元回归方程，该方程的形式如下所示。

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

应变量$\hat{y}$是一位客户属于一个特定的客户群体的一个概率估计。自变量x_1，x_2，⋯，x_p是客户的实际购买行为的度量，可能包括购买的具体物品、购买物品的数量、购买物品的金额、周几购买、一天的什么时候购买，等等。分析有助于确认出与预测客户群体最相关的自变量，对客户总体提供了更好的了解，能以更大的信心做进一步的分析。分析的重点在于了解顾客对商品开发、市场营销、直接营销方案等问题的态度，而这些问题与公司服务的客户群有极大的相关性。

在本章中，我们将介绍多元回归的概念，并且说明如何将第14章介绍的简单线性回归的概念扩展到多元回归的情形。另外，我们还将说明，对于多元回归问题如何使用计算机软件包的方法。在本章的最后一节，我们将通过一个例题介绍logistic回归，该例题说明了在市场营销研究中如何应用logistic回归方法。

在第14章中，我们介绍了简单线性回归问题，说明了它在建立描述两个变量之间关系的估计的回归方程中的应用。让我们回想一下，被回归方程预测的或者解释的变量称为应变量，用来预测或者解释应变量的变量称为自变量。在这一章中，我们将通过考虑两个或两个以上自变量的情形，来继续我们回归分析的学习。这一研究领域被称为**多元回归分析**（multiple regression analysis）。多元回归分析使我们能够考虑较多的因素，并且能得到比利用简单线性回归更好的预测结果。

15.1 多元回归模型

多元回归分析是研究应变量y如何依赖两个或两个以上自变量的问题。在一般情形下，我们将用p示自变量的数目。

15.1.1 回归模型和回归方程

我们在上一章介绍的回归模型和回归方程的概念对多元回归情形是同样适用的。描述应变量y如何依赖于自

⊖ 作者感谢dunnhumby公司解决方案高级副总裁Paul Hunter先生，他为“实践中的统计”提供了本案例。

变量 x_1，x_2，…，x_p 和一个误差项的方程被称为**多元回归模型**（multiple regression model）。我们首先假设多元回归模型的形式如下。

多元回归模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon \tag{15-1}$$

在多元回归模型中，β_0，β_1，β_2，…，β_p 是参数，误差项 ε（希腊字母）是随机变量。对这一模型细加观察后就会发现，y 是 x_1，x_2，…，x_p 的线性函数（$\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$ 部分）加上一个误差项 ε。误差项说明了包含在 y 里面但不能被 p 个自变量的线性关系解释的变异性。

在第 15.4 节中，我们将讨论多元回归模型和 ε 的假定。假定之一是 ε 的平均值或期望值是零。由于这一假定，y 的平均值或期望值 $E(y)$ 等于 $\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$。描述 y 的平均值如何依赖于 x_1，x_2，…，x_p 的方程被称为**多元回归方程**（multiple regression equation）。

多元回归方程

$$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p \tag{15-2}$$

15.1.2 估计的多元回归方程

如果参数 β_0，β_1，β_2，…，β_p 的值是已知的，在给定 x_1，x_2，…，x_p 的值时，我们能利用式（15-2）计算 y 的平均值。遗憾的是，这些参数的值通常都是未知的，我们必须利用样本数据去估计它们。我们利用一个简单随机样本计算样本统计量 b_0，b_1，b_2，…，b_p，将它们作为未知参数 β_0，β_1，β_2，…，β_p 的点估计量。利用这些样本统计量使我们得到了下面的**估计的多元回归方程**（estimated multiple regression equation）。

估计的多元回归方程

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p \tag{15-3}$$

式中，b_0，b_1，b_2，…，b_p 是 β_0，β_1，β_2，…，β_p 的估计值；$\hat{y}$ 代表应变量的预测值。

对于多元回归情形，估计步骤如图 15-1 所示[⊖]。

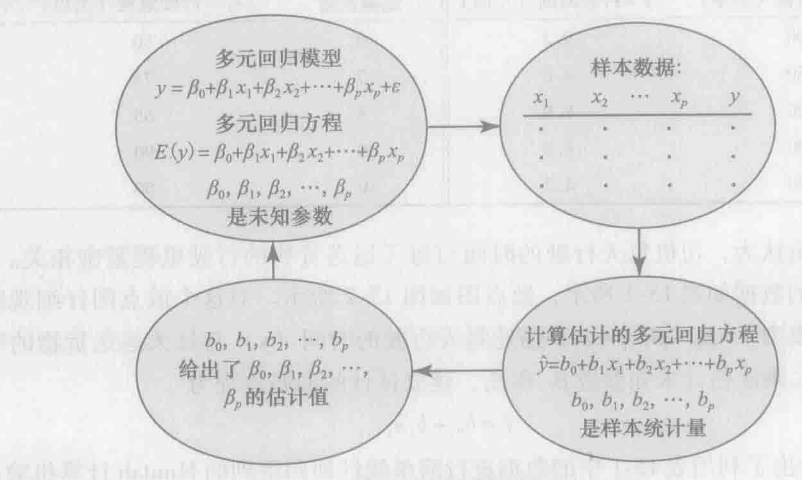


图 15-1　多元性回归模型的估计步骤

⊖ 在简单线性回归情形中，b_0 和 b_1 是用于估计未知参数 β_0 和 β_1 的样本统计量。在多元回归类似的统计推断过程中，我们用 b_0，b_1，b_2，…，b_p 表示用于估计未知参数 β_0，β_1，β_2，…，β_p 的样本统计量。

15.2 最小二乘法

在第14章，我们利用**最小二乘法**（least squares method）建立了估计的回归方程，这个方程最佳地近似了应变量和自变量之间的直线关系。现在我们利用同样的方法来建立估计的多元回归方程。我们再次叙述最小二乘准则如下。

最小二乘法准则

$$\min \sum (y_i - \hat{y}_i)^2 \tag{15-4}$$

式中，y_i 代表对于第 i 次观测，应变量的观测值；$\hat{y}_i$ 代表对于第 i 次观测，应变量的预测值。

我们能利用估计的多元回归方程

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

计算应变量的预测值。

正如式（15-4）所显示的那样，最小二乘法是利用样本数据，通过使残差［即应变量的观测值（y_i）与应变量的预测值（$\hat{y}_i$）之间的离差］的平方和达到最小的方法求得 b_0，b_1，b_2，…，b_p的值。

为了估计简单线性回归方程$\hat{y} = b_0 + b_1x$，在第14章里，我们介绍了计算最小二乘估计量 b_0 和 b_1的公式。对于样本容量相对比较小的数据集，我们能利用这些公式通过手算就能计算出 b_0 和 b_1。可是在多元回归情形中，计算回归系数 b_0，b_1，b_2，…，b_p的公式将涉及矩阵代数的知识，这已经超出了本书的范围。所以，在介绍多元回归时，我们将把注意力集中到如何使用计算机软件包来得到估计的回归方程以及其他的有关信息。重点将是如何解释计算机软件包的输出，而不是如何进行多元回归的计算。

15.2.1 一个例子：Butler 运输公司

作为多元回归分析的一个说明，我们将考虑 Butler 运输公司所面临的一个问题，它是一家位于南加利福尼亚地区的独立运输公司。Butler 运输公司的主要业务遍及它所在地区的货物运送。为了制定最佳的工作计划表，公司的管理人员希望估计他们的司机每天行驶的时间，如表15-1所示。

表15-1 Butler 运输公司的原始数据

运输任务	x_1 = 行驶里程（英里）	y = 行驶时间（小时）	运输任务	x_1 = 行驶里程（英里）	y = 行驶时间（小时）
1	100	9.3	6	80	6.2
2	50	4.8	7	75	7.4
3	100	8.9	8	65	6.0
4	100	6.5	9	90	7.6
5	50	4.2	10	90	6.1

最初，公司的管理人员认为，司机每天行驶的时间与每天运送货物的行驶里程紧密相关。由10项运输任务组成的简单随机样本提供的数据如表15-1所示，散点图如图15-2所示。对这个散点图仔细观察后，管理人员假设，能利用简单线性回归模型 $y = \beta_0 + \beta_1x_1 + \varepsilon$ 来描述每天行驶的时间（y）与每天运送货物的行驶里程（x_1）之间的关系。我们利用最小二乘法估计未知参数 β_0 和 β_1，建立估计的回归方程为

$$\hat{y} = b_0 + b_1x_1 \tag{15-5}$$

在图15-3中，[㊀]我们给出了利用表15-1中的数据进行简单线性回归得到的 Minitab 计算机输出。估计的回归方程是

$$\hat{y} = 1.27 + 0.0678x_1$$

㊀ 在 Minitab 计算机输出中，我们将变量名 Miles 和 Time 作为列标题输入到工作表中。于是 x_1 = Miles 和 y = Time。

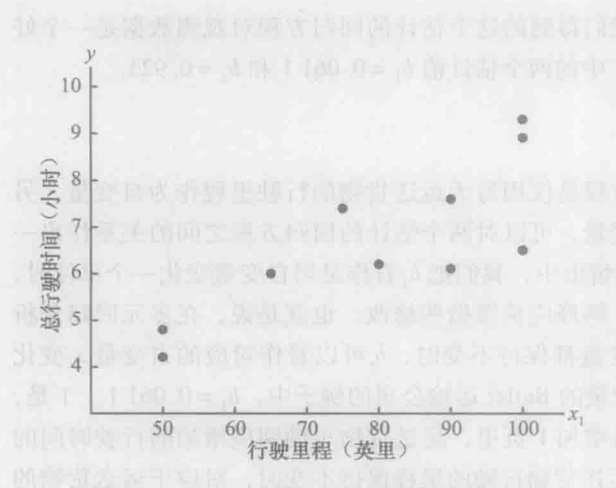


图 15-2　Butler 运输公司原始数据的散点图

回归方程是销售额
Time=1.27+0.067 8 Miles

预测量	系数	标准差	T	p
常数	1.274	1.401	0.91	0.390
Miles	0.067 83	0.017 06	3.98	0.004

S=1.001 79　R-Sq=66.4%　R-Sq (adj)=62.2%

方差分析

来源	自由度	平方和	均方	F	p
回归	1	15.871	15.871	15.81	0.004
误差	8	8.029	1.004		
总计	9	23.900			

图 15-3　含有一个自变量的 Butler 运输公司的 Minitab 输出

F 统计量的值为 15. 81，它所对应的 p- 值为 0. 004，在 $\alpha = 0.05$ 的显著性水平下表明，y 和 x_1 之间的关系是显著的；也就是说，由于 p- 值比 $\alpha = 0.05$ 小，因此我们应该拒绝原假设 H_0：$\beta_1 = 0$。注意，因为 t 统计量的值为 3. 98，与它相联系的 p- 值为 0. 004，所以我们也可以得到同样的结论。于是我们能够断定，每天行驶的时间与每天运送货物的行驶里程之间存在一个显著的关系；每天比较长的行驶时间与每天比较多的运送货物的行驶里程相对应。因为判定系数 $R^2 = 66.4\%$（用一个百分数表示），所以我们可以看出，运输车辆每天行驶时间变异性的 66. 4% 能被运送货物的行驶里程的线性影响解释。这一结果还算不错，但是管理人员或许希望考虑增加第二个自变量去解释应变量中剩余的变异性。

在试图确认另一个自变量时，管理人员觉得运送货物的次数也可能影响到行驶的总时间。Butler 运输公司增加了运送货物次数的数据，如表 15-2 所示。用每天运送货物的行驶里程（x_1）和运送货物的次数（x_2）作自变量，Minitab 计算机输出结果如图 15-4 所示。[⊖] 估计的回归方程是

$$\hat{y} = -0.869 + 0.061\,1x_1 + 0.923x_2 \tag{15-6}$$

表 15-2　用行驶里程（x_1）和运送货物次数（x_2）作自变量的 Butler 运输公司的数据

运输任务	x_1 = 行驶里程（英里）	x_2 = 运送货物的次数	y = 行驶时间（小时）
1	100	4	9. 3
2	50	3	4. 8
3	100	4	8. 9
4	100	2	6. 5
5	50	2	4. 2
6	80	2	6. 2
7	75	3	7. 4
8	65	4	6. 0
9	90	3	7. 6
10	90	2	6. 1

回归方程是销售额
Time=−0.869+0.061 1 Miles+0.923 Deliveries

预测量	系数	标准差	T	p
常数	−0.868 7	0.951 5	−0.91	0.392
Miles	0.061 135	0.009 888	6.18	0.000
Deliveries	0.923 4	0.221 1	4.18	0.004

S=0.573 142　R-Sq=90.4%　R-Sq (adj)=87.6%

方差分析

来源	自由度	平方和	均方	F	p
回归	2	21.601	10.800	32.88	0.000
误差	7	2.299	0.328		
总计	9	23.900			

图 15-4　具有两个自变量的 Butler 运输公司的 Minitab 输出[⊜]

⊖ 生成如图 15-4 所示的 Minitab 输出的必要步骤在附录 15A 中给出。

⊜ 在 Minitab 计算机输出中，我们将变量名 Miles、Deliveries 和 Time 作为列标题输入到工作表中。于是 x_1 = Miles、x_2 = Deliveries 和 y = Time。

在下一节，我们将讨论如何利用多元判定系数来度量，我们得到的这个估计的回归方程对观测数据是一个好的拟合。在此之前，让我们先更仔细地考察一下在式（15-6）中的两个估计值 $b_1=0.0611$ 和 $b_2=0.923$。

15.2.2 关于回归系数解释的注释

现在我们得到了两个估计的回归方程，一个估计的回归方程是仅用每天运送货物的行驶里程作为自变量，另一个估计的回归方程包含了用运送货物的次数作为第二个自变量，可以对两个估计的回归方程之间的关系作出一些解释。在这两种情形下，b_1 的值是不同的。在简单线性回归情形中，我们把 b_1 看作是当自变量变化一个单位时，应变量 y 变化程度的一个估计。在多元回归分析情形中，这一解释应稍微做些修改。也就是说，在多元回归分析情形中，我们对每一个回归系数有如下解释：当所有其他自变量都保持不变时，b_j 可以看作对应的自变量 x_i 变化一个单位时，应变量 y 变化程度的一个估计。在包括两个自变量的 Butler 运输公司的例子中，$b_1=0.0611$。于是，当运送货物的次数保持不变时，对应于运送货物行驶的里程每增加 1 英里，运送货物车辆期望增加的行驶时间的估计值是 0.0611 个小时。类似地，因为 $b_2=0.923$，所以当运送货物行驶的里程保持不变时，对应于运送货物的次数每增加 1 次，运送货物车辆期望增加的行驶时间的估计值是 0.923 个小时。

练 习

请注意：在设计本节和后面几节，解答包括数据在内的习题时，我们是借助于计算机软件包完成的。

方法

☆2. 对于应变量 y 和两个自变量 x_1 和 x_2，考虑下面的数据。

x_1	x_2	y	x_1	x_2	y
30	12	94	51	19	175
47	10	108	74	7	170
25	17	112	36	12	117
51	16	178	59	13	142
40	5	94	76	16	211

a. 利用这些数据，建立 y 关于 x_1 的估计的回归方程。如果 $x_1=45$，预测 y 的值。

b. 利用这些数据，建立 y 关于 x_2 的估计的回归方程。如果 $x_2=15$，预测 y 的值。

c. 利用这些数据，建立 y 关于 x_1 和 x_2 的估计的回归方程。如果 $x_1=45$，$x_2=15$，预测 y 的值。

应用

4. 一家鞋店建立了销售收入关于存货投资和广告费用的估计的回归方程如下。

$$\hat{y}=25+10x_1+8x_2$$

式中，x_1 为存货投资（1 000 美元）；x_2 为广告费用（1 000 美元）；y 为销售收入（1 000 美元）。

a. 如果这家鞋店有存货投资 15 000 美元，广告预算费用为 10 000 美元，预测该家鞋店的销售收入。

b. 对估计的回归方程中的 b_1 和 b_2 作出解释。

6. 美国国家橄榄球联盟记录了每名球员和每支球队各种表现的数据。为了调查传球在球队获胜中发挥的重要作用，由 16 支美国国家橄榄球联盟的球队组成一个随机样本，每支球队所属的联合会、在 2011 年赛季每次进攻的平均传球码数（Yds/Att）、每次进攻被守方拦截的次数（Int/Att）和比赛的获胜率（Win%）数据如下表所示（NFL website，2012 年 2 月 12 日）。

球队名称	所属联合会	Yds/Att	Int/Att	Win（%）
Arizona Cardinals	NFC	6.5	0.042	50.0
Atlanta Falcons	NFC	7.1	0.022	62.5
Carolina Panthers	NFC	7.4	0.033	37.5
Cincinnati Bengals	AFC	6.2	0.026	56.3
Detroit Lions	NFC	7.2	0.024	62.5
Green Bay Packers	NFC	8.9	0.014	93.8
Houston Texans	AFC	7.5	0.019	62.5
Indianapolis Colts	AFC	5.6	0.026	12.5
Jacksonville Jaguars	AFC	4.6	0.032	31.3
Minnesota Vikings	NFC	5.8	0.033	18.8
New England Patriots	AFC	8.3	0.020	81.3
NewOrleans Saints	NFC	8.1	0.021	81.3
Oakland Raiders	AFC	7.6	0.044	50.0
San Francisco 49ers	NFC	6.5	0.011	81.3
Tennessee Titans	AFC	6.7	0.024	56.3
Washington Redskins	NFC	6.4	0.041	31.3

注：AFC 为美国美式橄榄球联合会，NFC 为国家美式橄榄球联合会。

a. 建立估计的回归方程，使这个方程在球队每次进攻的平均传球码数已知时，能用来预测比赛的获胜率。

b. 建立估计的回归方程，使这个方程在球队每次进攻被守方拦截的次数已知时，能用来预测比赛的获胜率。

c. 建立估计的回归方程，使这个方程在球队每次进攻的平均传球码数、每次进攻被守方拦截的次数已知时，能用来预测比赛的获胜率。

d. 在2011年赛季，Kansas City Chiefs队在比赛中每次进攻的平均传球码数是6.2，每次进攻被守方拦截的次数是0.036。利用在（c）中建立的估计的回归方程，预测该队的获胜率。（注：在2011年赛季，Kansas City Chiefs队的记录是7胜9负。）将你的预测结果与Kansas City Chiefs队实际比赛的获胜率进行比较。

8. *Conde Nast Traveler* 杂志金牌榜对2012年排名前20位的小型豪华游轮评定了等级（*Conde Nast Traveler* website，2012年3月）根据 *Conde Nast Traveler* 杂志的年度"读者选择调查"（Readers' Choice Survey）结果，得到的每一艘游轮的等级得分如下表所示。每个得分表示受访者按照一些标准，将一艘游轮评定为优秀或很好的比例，包括岸上观光和食品/餐饮。表中还给出了总得分和进行等级排序的游轮。*Seabourn Odyssey* 是排名最高的游轮，总得分为94.4，其中得分最高的是食品/餐饮项，为97.8。

游轮名称	总得分	岸上观光得分	食品/餐饮得分
Seabourn Odyssey	94.4	90.9	97.8
Seabourn Pride	93.0	84.2	96.7
National Geographic Endeavor	92.9	100.0	88.5
Seabourn Sojourn	91.3	94.8	97.1
Paul Gauguin	90.5	87.9	91.2
Seabourn Legend	90.3	82.1	98.8
Seabourn Spirit	90.2	86.3	92.0
Silver Explorer	89.9	92.6	88.9
Silver Spirit	89.4	85.9	90.8
Seven Seas Navigator	89.2	83.3	90.5
Silver Whisperer	89.2	82.0	88.6
National Geographic Explorer	89.1	93.1	89.7
Silver Cloud	88.7	78.3	91.3
CelebrityXpedition	87.2	91.7	73.6
Silver Shadow	87.2	75.0	89.7
Silver Wind	86.6	78.1	91.6
Sea Dream Ⅱ	86.2	77.4	90.9
Wind Star	86.1	76.5	91.5
Wind Surf	86.1	72.3	89.3
Wind Spirit	85.2	77.4	91.9

a. 建立估计的回归方程，使这个方程在岸上观光得分已知时，能用来预测游轮的总得分。

b. 考虑增加食品/餐饮得分为自变量。建立估计的回归方程，使这个方程在岸上观光得分和食品/餐饮得分已知时，能用来预测游轮的总得分。

c. 对于一艘岸上观光得分为80分，食品/餐饮得分为90的豪华游轮，预测该游轮的总得分。

10. 美国职业棒球大联盟由美国联盟和国家联盟的球队组成。它们收集了球队和球员的各种各样的统计数据。经常用来评价投球表现的一些统计量如下：

ERA（投手责任失分率）：每九局比赛中，由于投手投球的责任造成对方得分的平均次数。投手责任失分是由于投手的因素所造成的失分，而由于防守失误使对方跑垒者上垒所失的分数不记为投手责任失分。

SO/IP：每局比赛投手使对方击球员三击未中出局的平均次数。

HR/IP：每局比赛投手被对方击球员本垒打的平均次数。

R/IP：每局比赛的投手平均责任失分。

由2011赛季美国联盟的20名投手组成一个随机样本，上述这些统计量的数值如下表所示（MLB website，2012年3月1日）。

投手	球队名称	胜	负	ERA	SO/IP	HR/IP	R/IP
Verlander, J	DET	24	5	2.40	1.00	0.10	0.29
Beckett, J	BOS	13	7	2.89	0.91	0.11	0.34
Wilson, C	TEX	16	7	2.94	0.92	0.07	0.40
Sabathia, C	NYY	19	8	3.00	0.97	0.07	0.37
Haren, D	LAA	16	10	3.17	0.81	0.08	0.38
McCarthy, B	OAK	9	9	3.32	0.72	0.06	0.43
Santana, E	LAA	11	12	3.38	0.78	0.11	0.42
Lester, J	BOS	15	9	3.47	0.95	0.10	0.40
Hernandez, F	SEA	14	14	3.47	0.95	0.08	0.42
Buehrle, M	CWS	13	9	3.59	0.53	0.10	0.45

（续）

投手	球队名称	胜	负	ERA	SO/IP	HR/IP	R/IP
Pineda, M	SEA	9	10	3.74	1.01	0.11	0.44
Colon, B	NYY	8	10	4.00	0.82	0.13	0.52
Tomlin, J	CLE	12	7	4.25	0.54	0.15	0.48
Pavano, C	MIN	9	13	4.30	0.46	0.10	0.55
Danks, J	CWS	8	12	4.33	0.79	0.11	0.52
Guthrie, J	BAL	9	17	4.33	0.63	0.13	0.54
Lewis, C	TEX	14	10	4.40	0.84	0.17	0.51
Scherzer, M	DET	15	9	4.43	0.89	0.15	0.52
Davis, W	TB	11	10	4.45	0.57	0.13	0.52
Porcello, R	DET	14	9	4.75	0.57	0.10	0.57

a. 建立一个估计的回归方程，使这个方程在投手每局比赛使对方击球员三击未中出局的平均次数已知时，能用来预测每局比赛的投手平均责任失分。

b. 建立一个估计的回归方程，使这个方程在投手每局比赛被对方击球员本垒打的平均次数已知时，能用来预测每局比赛的投手平均责任失分。

c. 建立估计的回归方程，使这个方程在投手每局比赛使对方击球员三击未中出局的平均次数和被对方击球员本垒打的平均次数已知时，能用来预测每局比赛的投手平均责任失分。

d. A. J. Burnett 是 New York Yankees 队的一名投手，他在每局比赛中使对方击球员三击未中出局的平均次数是 0.91，被对方击球员本垒打的平均次数是 0.16。利用在（c）中建立的估计的回归方程，来预测该名投手在每局比赛中的平均责任失分。（注意：A. J. Burnett 在每局比赛中的实际平均责任失分为 0.6）

e. 假设有人建议：在（c）中还使用 ERA（投手责任失分率）作为另一个自变量。你怎么看待这个建议？

15.3　多元判定系数

在简单线性回归情形中，我们已经说明了总的平方和能被分解为两部分：回归平方和及误差平方和。同样的方法也能应用到多元回归平方和的情形。

SST，SSR 和 SSE 之间的关系

$$\text{SST} = \text{SSR} + \text{SSE} \tag{15-7}$$

式中，$\text{SST} = \sum(y_i - \bar{y})^2$ 代表总的平方和；$\text{SSR} = \sum(\hat{y}_i - \bar{y})^2$ 代表回归平方和；$\text{SSE} = \sum(y_i - \hat{y}_i)^2$ 代表误差平方和。

由于这三个平方和的计算非常困难，因此我们依靠计算机软件包得到这三个平方和的数值。对于具有两个自变量的 Butler 运输公司问题，图 15-4 的 Minitab 输出的方差分析部分给出了这三个平方和的数值：SST = 23.900，SSR = 21.601 和 SSE = 2.299。当仅有一个自变量（每天运送货物的行驶里程）时，图 15-3 的 Minitab 输出给出了 SST = 23.900，SSR = 15.871 和 SSE = 8.029。在这两种情形下，SST 的值是相同的，因为它不依赖于 $\hat{y}$。但是当第二个自变量（运送货物的次数）进入模型后，SSR 增加而 SSE 减少。这就意味着，估计的多元回归方程对观测数据给出了一个更好的拟合。

在第 14 章里，我们利用判定系数 $r^2 = \text{SSR}/\text{SST}$ 来度量估计的回归方程的拟合优度。我们将同样的概念应用到多元回归的情形。术语**多元判定系数**（multiple coefficient of determination）表示的是对估计的多元回归方程拟合优度的度量。我们用 R^2 表示多元判定系数，它的计算公式如下。

多元判定系数

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{15-8}$$

我们能把多元判定系数 R^2 理解为应变量 y 中的变异性能被估计的多元回归方程解释的比例。将这个比例乘以 100，我们就能把多元判定系数 R^2 理解为应变量 y 中的变异性能被估计的多元回归方程解释的百

分比。

对于有两个自变量的 Butler 运输公司的例子，因为 SSR = 21.601，SST = 23.900，我们有

$$R^2 = \frac{21.601}{23.900} = 0.904$$

所以，在运输车辆行驶时间 y 中变异性的 90.4%，能用运送货物的行驶里程和运送货物的次数作自变量的估计的多元回归方程解释。在图 15-4 中我们看到，Minitab 输出还给出了多元判定系数（表示为一个百分数），被表示为 R－Sq = 90.4%。

对于仅有一个自变量，即每天运送货物的行驶里程（x_1）的估计的多元回归方程，图 15-3 给出了 R－Sq 的值是 66.4%。于是，当运送货物的次数作为第二个自变量进入模型后，运输车辆行驶时间 y 的变异性中能被估计的多元回归方程解释的百分比由 66.4% 增加到 90.4%。在一般情形下，R^2 总是随着新的自变量不断地进入模型而增加[㊀]。

由于增加自变量将影响到应变量中的变异性被估计的回归方程解释的百分比，为了避免高估这一影响，许多分析学家提出用自变量的数目去修正 R^2 的值。用 n 表示观测值的数目，p 表示自变量的数目，**修正多元判定系数**（adjust multiple coefficient of determination）的计算公式如下。[㊁]

修正多元判定系数

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{15-9}$$

对于 Butler 运输公司的例子，$n = 10$，$p = 2$，我们有

$$R_a^2 = 1 - (1 - 0.904)\frac{10-1}{10-2-1} = 0.88$$

于是，对于有两个自变量的多元判定系数进行修正后，我们得到的修正多元判定系数是 0.88。在图 15-4 中，这一数值（表示为一个百分数）由 Minitab 输出给出，它被表示为 R－Sq(adj) = 87.6%；这一数值与我们的计算结果不一样，这是因为我们在计算时，用的是 R^2 四舍五入的数值。

注释和评论

如果 R^2 的数值比较小，而模型所包含的自变量的数目比较大时，修正多元判定系数可能取负值；在这种情形下，Minitab 将把修正多元判定系数的数值调整为 0。

练习

方法

☆12. 在第 2 题中，给出了应变量 y 和两个自变量 x_1，x_2 的 10 次观测结果；对这些数据我们计算出 SST = 15 182.9，SSR = 14 052.2。

a. 计算 R^2。

b. 计算 R_a^2。

c. 估计的回归方程是否解释了数据中的大部分变异性？请作出解释。

应用

14. 在第 4 题中，我们给出了销售收入关于存货投资和广告费用的估计的回归方程如下。

$$\hat{y} = 25 + 10x_1 + 8x_2$$

通过对 10 家鞋店的一次统计调查，得到了用于建立模型的数据；对这些数据，我们计算出 SST = 16 000，SSR = 12 000。

a. 对于已给的估计的回归方程，计算 R^2。

㊀ 增加自变量将使预测误差变得比较小，从而减少误差平方和 SSE。因为 SSR = SST－SSE，当 SSE 变得比较小时，SSR 就变得比较大，从而使得 R^2 = SSR/SST 增加。

㊁ 如果模型增加一个自变量，即使这个增加的自变量在统计上并不显著，R^2 也将变得比较大。修正多元判定系数抵消了模型中自变量个数的影响。

b. 计算 R_a^2。

c. 估计的回归方程是否解释了数据中的大部分变异性？请作出解释。

16. 在第 6 题中，对于由 16 支美国国家橄榄球联盟（NFL）的球队组成一个随机样本，我们给出了这些球队在 2011 年赛季每次进攻的平均传球码数（Yds/Att）、每次进攻被守方拦截的次数（Int/Att）和比赛的获胜率（Win%）数据（NFL website，2012 年 2 月 12 日）。

a. 如果仅仅利用球队每次进攻的平均传球码数（Yds/Att）作为自变量来预测球队比赛的获胜率，这样得到的估计的回归方程，能否给出一个好的拟合？

b. 如果利用球队每次进攻的平均传球码数（Yds/Att）、每次进攻被守方拦截的次数（Int/Att）两个自变量来预测球队比赛的获胜率，讨论这样得到的估计的回归方程的优点。

18. 在第 10 题中，我们已经得到了由 2011 赛季美国联盟的 20 名投手组成的一个随机样本，以及有关这些投手的统计量（MLB website，2012 年 3 月 1 日）。

a. 在第 10 题的（c）中，我们已经建立了一个估计的回归方程，该方程在投手每局比赛使对方击球员三击未中出局的平均次数和被对方击球员本垒打的平均次数已知时，能用来预测每局比赛的投手平均责任失分。R^2 和 R_a^2 的数值是多少？

b. 估计的回归方程对观测数据的拟合好吗？请作出解释。

c. 假定在（c）中，用投手责任失分率（ERA）作为应变量，来替代每局比赛的投手平均责任失分。利用 ERA 得到的估计的回归方程对观测数据的拟合好吗？请作出解释。

15.4 模型的假定

在第 15.1 节中，我们引入了式（15-10）形式的多元回归模型。

多元回归模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon \tag{15-10}$$

在多元回归模型中，关于误差项 ε 的假定与简单线性回归模型的那些假定相似。

关于多元回归模型 $y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon$ 的误差项 ε 的假定

1. 误差项 ε 是一个平均值或期望值为零的随机变量；即，$E(\varepsilon)=0$。

这就意味着：对于给定的 x_1，x_2，…，x_p 的值，y 的期望值或平均值是

$$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p \tag{15-11}$$

式（15-11）就是我们在第 15.1 节中介绍的多元回归方程。在这个方程中，$E(y)$ 表示对于给定的 x_1，x_2，…，x_p 的值，y 的所有可能出现的值的期望值或平均值。

2. 对于自变量 x_1，x_2，…，x_p 所有的值，ε 的方差都是相同的。我们用 σ^2 表示 ε 的方差。

这就意味着：对于自变量 x_1，x_2，…，x_p 所有的值，y 关于回归线的方差也都等于 σ^2。

3. ε 的值是相互独立的。

这就意味着：对于自变量 x_1，x_2，…，x_p 的一组特定值，它所对应的误差项 ε 的值与自变量 x_1，x_2，…，x_p 的任意一组其他值所对应的误差项 ε 的值不相关。

4. 误差项 ε 是一个服从正态分布的随机变量，它表示了 y 的值和由 $\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$ 给出的 y 的期望值之间的离差。

这就意味着：因为 β_0，β_1，β_2，…，β_p 是常数，对于给定的 x_1，x_2，…，x_p 的值，应变量 y 也是一个服从正态分布的随机变量。

为了对由式（15-11）给出的关系式的形式有更深入的了解，我们考虑下面含有两个自变量的多元回归方程。

$$E(y)=\beta_0+\beta_1x_1+\beta_2x_2$$

这个方程的图形是三维空间的一个平面。图15-5给出了这样一个图形的例子。注意，在图中，当 $x_1=x_1^*$，$x_2=x_2^*$ 时，ε 是 y 的实际观测值和 y 的期望值 $E(y)$ 之间的差。

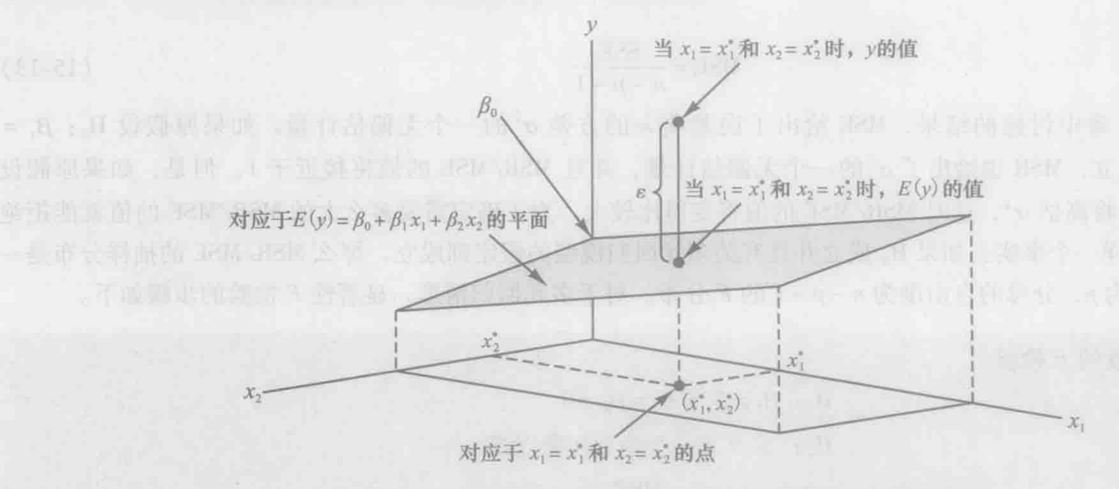


图15-5 具有两个自变量的多元回归分析问题的回归方程图形

在回归分析中，我们经常用术语响应变量代替术语应变量。此外，因为多元回归方程生成一个平面或是一个曲面，所以我们把它的图形称为响应曲面。

15.5 显著性检验

在这一节我们将说明，如何对多元回归关系进行显著性检验。在简单线性回归情形中，我们应用 t 检验和 F 检验进行显著性检验。在简单线性回归情形，这两个检验给出了同样的结论；也就是，如果原假设被拒绝，我们的结论是 $\beta_1\neq0$。而在多元回归情形中，进行 t 检验和 F 检验的目的是不同的。

（1）F 检验用于确定在应变量和所有自变量之间是否存在一个显著性的关系；我们把 F 检验称为总体的显著性检验。

（2）如果 F 检验已经表明了模型总体的显著性，那么 t 检验用于来确定每一个单个的自变量是否为一个显著的自变量。对模型中每一个单个的自变量，都要单独地进行 t 检验；我们把每一个这样的 t 检验都称为单个的显著性检验。

在以下内容中，我们将对 F 检验和 t 检验加以解释，并将这两个检验应用于Butler运输公司的例子上。

15.5.1 F检验

在第15.4节中定义的多元回归模型是

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon$$

F 检验的假设与多元回归模型的参数有关。

$$H_0: \beta_1=\beta_2=\cdots=\beta_p=0$$

$$H_a: \text{至少有一个参数不等于零}$$

如果原假设 H_0 被拒绝，那么我们就有足够的统计证据断定，至少有一个参数不等于零，并且应变量 y 和所有自变量 x_1，x_2，…，x_p 之间的关系在总体上是显著的。如果原假设 H_0 没有被拒绝，我们就没有足够的理由认为，y 和 x_1，x_2，…，x_p 之间存在一个显著的关系。

在叙述 F 检验的步骤之前，我们需要回忆一下均方的概念。均方是一个平方和除以它所对应的自由度。在多

元回归情形中，总的平方和有 $n-1$ 个自由度，回归平方和SSR有 p 个自由度，误差平方和SSE有 $n-p-1$ 个自由度。因此，均方回归（MSR）是 SSR/p，均方误差（MSE）是 SSE/$(n-p-1)$。

$$\text{MSR}=\frac{\text{SSR}}{p} \tag{15-12}$$

和

$$\text{MSE}=\frac{\text{SSE}}{n-p-1} \tag{15-13}$$

根据在第 14 章中讨论的结果，MSE 给出了误差项 ε 的方差 σ^2 的一个无偏估计量。如果原假设 H_0：$\beta_1=\beta_2=\cdots=\beta_p=0$ 成立，MSR 也给出了 σ^2 的一个无偏估计量，并且 MSR/MSE 的值将接近于 1。但是，如果原假设 H_0 被拒绝，MSR 将高估 σ^2，这时 MSR/MSE 的值将变得比较大。为了确定需要多么大的 MSR/MSE 的值就能拒绝 H_0，我们利用这样一个事实：如果 H_0 成立并且有关多元回归模型的假定都成立，那么 MSR/MSE 的抽样分布是一个分子的自由度为 p，分母的自由度为 $n-p-1$ 的 F 分布。对于多元回归情形，显著性 F 检验的步骤如下。

总体显著性的 F 检验

$$H_0:\ \beta_1=\beta_2=\cdots=\beta_p=0$$

$$H_a:\ \text{至少有一个参数不等于零}$$

检验的统计量

$$F=\frac{\text{MSR}}{\text{MSE}} \tag{15-14}$$

拒绝法则

p-值法：如果 p-值 $\leqslant\alpha$，则拒绝 H_0

临界值法：如果 $F\geqslant F_\alpha$，则拒绝 H_0

式中，F_α 代表分子自由度为 p，分母自由度为 $n-p-1$ 时，使 F 分布的上侧面积为 α 的 F 值。

让我们将 F 检验应用于 Butler 运输公司的多元回归问题。因为有两个自变量，所以我们把检验的假设写成如下形式。

$$H_0:\ \beta_1=\beta_2=0$$

$$H_a:\ \beta_1,\ \beta_2 \text{中至少有一个不等于零}$$

图 15-6 是用每天运送货物的行驶里程（x_1）和运送货物的次数（x_2）作为自变量的多元回归模型的 Minitab 计算机输出。在 Minitab 输出的方差分析部分，我们看到 MSR = 10.8，MSE = 0.328。利用式（15-14），我们得到检验的统计量

$$F=\frac{10.8}{0.328}=32.9$$

注意，Minitab 输出的 F 值是 32.88，与我们算出的数值不同，这是由我们在计算过程中使用了 MSR 和 MSE 的四舍五入近似值引起的。对于 $\alpha=0.01$ 的显著性水平，在图 15-6 方差分析部分的最后一列我们看到 p-值 = 0.000，因为 p-值小于 0.01，所以应该拒绝 H_0。换一种叙述方式，我们还可以从附录 B 的表 B-4 中，查到分子自由度为 2，分母自由度为 7 的 F 分布上侧分位数 $F_{0.01}=9.55$。因为 32.9 > 9.55，所以我们应该拒绝 H_0：$\beta_1=\beta_2=0$，并且得出结论：在每天行驶的时间 y 和每天运送货物的行驶里程 x_1、运送货物的次数 x_2 这两个自变量之间存在一个显著关系。

回归方程是销售额
Time=-0.869+0.061 1 Miles+0.923 Deliveries

预测量	系数	标准差	T	p
常数	-0.868 7	0.951 5	-0.91	0.392
Miles	0.061 135	0.009 888	6.18	0.000
Deliveries	0.923 4	0.221 1	4.18	0.004

S=0.573 142 R-Sq=90.4% R-Sq(adj)=87.6%

方差分析

来源	自由度	平方和	均方	F	p
回归	2	21.601	10.800	32.88	0.000
误差	7	2.299	0.328		
总计	9	23.900			

图 15-6 以行驶里程（x_1）和运送货物次数（x_2）为自变量的

前面已经提到，均方误差 MSE 给出了误差项 ε 的方差 σ^2 的一个无偏估计量。从图 15-6 中我们看到，σ^2 的估计量是 MSE =

0.328。MSE的平方根是误差项 ε 标准差的估计。就像在第14.5节中所定义的那样，我们把这个标准差称为估计的标准误差，并且用 s 表示。因此我们有 $s=\sqrt{MSE}=\sqrt{0.328}=0.573$。注意，在图15-6的Minitab计算机输出中，列出了估计的标准误差的值。

表15-3是一般的方差分析表（ANOVA），它给出了多元回归模型 F 检验的结果。F 检验统计量的数值在表的最后一列，我们把它和分子自由度为 p，分母自由度为 $n-p-1$ 的 α 水平的 F 分布上侧分位数 F_α 进行比较，从而得出假设检验的结论。再次审视图15-6中的Butler运输公司的Minitab输出，我们很快发现Minitab的方差分析表里包含了这些信息。此外，Minitab输出还给出了对应于 F 检验统计量的 p-值。

表15-3 具有 p 个自变量的多元回归模型的ANOVA表

方差来源	平方和	自由度	均方	F
回归	SSR	p	$MSR=\frac{SSR}{p}$	$F=\frac{MSR}{MSE}$
误差	SSE	$n-p-1$	$MSE=\frac{SSE}{n-p-1}$	
总和	SST	$n-1$		

15.5.2 t 检验

如果 F 检验显示了多元回归关系在总体上是显著的，那么 t 检验就能帮助我们确定每一个单个参数的显著性问题。对于单个参数的显著性检验，t 检验的步骤如下。

单个参数显著性的 t 检验

对于任一个参数 β_i

$$H_0: \beta_i=0$$

$$H_a: \beta_i\neq 0$$

检验的统计量

$$t=\frac{b_i}{s_{b_i}} \tag{15-15}$$

拒绝法则

p-值法：如果 p-值 $\leqslant\alpha$，则拒绝 H_0

临界值法：如果 $t\leqslant -t_{\alpha/2}$ 或者 $t\geqslant t_{\alpha/2}$，则拒绝 H_0

式中，$t_{\alpha/2}$ 是自由度为 $n-p-1$ 时，使 t 分布的上侧面积为 $\alpha/2$ 的 t 值。

在检验统计量的公式里，s_{b_i} 是 b_i 标准差的估计，s_{b_i} 的值是由计算机软件包提供的。

让我们将 t 检验用于Butler运输公司的多元回归问题。对于 t 统计量比值的计算，我们可以参考图15-6中的Minitab输出的有关部分。b_1，b_2，s_{b_1} 和 s_{b_2} 的值如下。

$$b_1=0.061\,135 \quad s_{b_1}=0.009\,888$$

$$b_2=0.923\,4 \quad s_{b_2}=0.221\,1$$

对于参数 β_1 和 β_2 的假设检验，我们利用式（15-15），就能得到检验的统计量。

$$t=0.061\,135/0.009\,888=6.18$$

$$t=0.923\,4/0.221\,1=4.18$$

注意，这两个 t 值和对应的 p-值都由图15-6的Minitab输出给出。对于显著性水平 $\alpha=0.01$，Minitab输出给出的 p-值分别为0.000和0.004，所以我们应该拒绝 H_0：$\beta_1=0$ 和 H_0：$\beta_2=0$。因此，这两个参数在统计上都是显著的。换一种叙述方式，对于显著性水平 $\alpha=0.01$，我们还可以从附录B的表B-2中，查出自由度为 $n-p-1=10-2-1=7$ 的 t 分布的双侧分位数 $t_{0.005}=3.499$。因为 $6.18>3.499$，所以我们应该拒绝 H_0：$\beta_1=0$。类似地，因为 $4.18>3.499$，所以我们应该拒绝 H_0：$\beta_2=0$。

15.5.3 多重共线性

在回归分析中，我们曾经使用自变量这个术语来表示用于预测或者解释应变量值的那些变量。但是这个术语并不意味着，这些自变量本身在某种统计意义上是独立的。相反，在多元回归问题中，大部分自变量在某种程度上是相互关联的。例如，在涉及两个自变量 x_1（每天运送货物的行驶里程）和 x_2（运送货物的次数）的 Butler 运输公司的例子中，我们可以把行驶里程看作应变量，把运送货物的次数看作自变量，来确定这两个变量本身是否是相关的。为此我们计算它们的样本相关系数 $r_{x_1x_2}$，然后确定这两个变量相关的程度。样本相关系数计算的结果是 $r_{x_1x_2}=0.16$。于是，我们发现在这两个自变量之间存在某种程度的线性联系。在多元回归分析中，我们把自变量之间的相关性称为**多重共线性**（multicollinearity）。

为了对多重共线性的潜在影响给出一个更好的全面描述，我们考虑对 Butler 运输公司的例子做一些修改。我们不用 x_2 表示运送货物的次数，而用 x_2 表示消耗汽油的加仑量。显然，x_1（每天运送货物的行驶里程）和 x_2 是相关的，即我们认为消耗汽油的加仑数量依赖于行驶的里程。因此，从逻辑上我们能认定 x_1 和 x_2 是高度相关的自变量。

假定，我们已经建立了估计的回归方程 $\hat{y}=b_0+b_1x_1+b_2x_2$，并且 F 检验显示了多元回归关系在总体上是显著的。然后，假定我们进行关于 β_1 的 t 检验来断定 $\beta_1\neq0$ 是否成立。如果我们不能拒绝 H_0：$\beta_1=0$，这个结果是否意味着每天运送货物行驶的时间不依赖于行驶的里程呢？不一定。它很可能意味着，由于 x_2 已经在模型里，所以 x_1 对决定 y 的数值已经不再有显著的贡献了。在我们的例子里，这一解释是能够讲通的，因为如果我们已经知道了消耗汽油的数量，在预测 y 的数值时，我们就不需要再增加更多的有关行驶里程的有用信息了。类似地，通过 t 检验可能导致我们得出 $\beta_2=0$ 结论，由于 x_1 已经在模型里，所以我们也不再需要增加更多的有关消耗汽油数量的信息了。

综上所述，在对单个参数的显著性进行 t 检验时，由于多重共线性而带来的困难是：当多元回归方程总体显著性的 F 检验表明有一个显著的关系时，我们可能得出单个参数没有一个是显著地不同于零的结论。只有当自变量之间的相关性非常小的时候，才有可能回避这个问题。

为了确定多重共线性是否对模型的估计产生十分严重的后果，统计学家已经提出了一些不同的检验方法。对于任何一个含有两个自变量的多元回归模型，如果它们的样本相关系数的绝对值大于 0.7，那么依照经验的检验方法，多重共线性有可能成为一个潜在的问题。[1]还有一些更精确的检验方法，但这些方法已经超出了本教科书讨论的范围。

如果可能的话，我们应尽一切努力避免在模型中包含高度相关的自变量。但是，在实际上我们很少有可能绝对坚持这一策略。决策者已经得到告诫，当他们有理由相信存在严重的多重共线性时，识别单个的自变量对应变量的影响将是一件非常困难的事情。[2]

注释和评论

通常，在对某个问题进行研究时，多重共线性对我们进行的回归分析，或者对计算机输出结果的解释都不会产生影响。但是，当存在严重的多重共线性时——也就是，当两个或两个以上的自变量相互高度相关时——如果我们对单个参数进行 t 检验，并对 t 检验的结果进行解释时，就可能出现困难。除了在本节说明的问题外，严重的多重共线性可能导致最小二乘估计出现错误的符号。也就是说，在模拟研究中，研究者设计了基本的回归模型，接着利用最小二乘法求出参数 β_0，β_1，β_2，…，的估计值，但是在高度多重共线性的条件下，最小二乘估计值可能与被估参数有一个完全相反的符号。例如，真实的 β_2 可能是 +10，而它的估计值 b_2 却是 −2。于是，如果存在高度的多重共线性，我们应能对个别系数的符号作出一些约定。

---

[1] 对于有两个自变量的情形，可能产生多重共线性潜在影响的一个比较粗略的估计方法是：如果它们的样本相关系数大于 0.7 或者小于 −0.7。

[2] 当自变量是高度相关时，不可能确定任一特定的自变量对应变量的单独影响。

练习

方法

20. 参阅在第2题中给出的数据。利用这些数据我们建立的估计的回归方程是

$$\hat{y} = -18.37 + 2.01x_1 + 4.74x_2$$

这里，SST = 15 182.9，SSR = 14 052.2，$s_{b_1} = 0.2471$ 和 $s_{b_2} = 0.9484$。

a. 在 $\alpha = 0.05$ 的显著性水平下，检验 x_1，x_2 和 y 之间的显著性关系。

b. 在 $\alpha = 0.05$ 的显著性水平下，β_1 是显著的吗？

c. 在 $\alpha = 0.05$ 的显著性水平下，β_2 是显著的吗？

应用

22. 在第4题中，我们给出了销售收入关于存货投资和广告费用的估计的回归方程如下。

$$\hat{y} = 25 + 10x_1 + 8x_2$$

通过对10家鞋店的一次统计调查，得到了用于建立模型的数据；对这些数据，我们计算出 SST = 16 000，SSR = 12 000。

a. 计算 SSE，MSE 和 MSR。

b. 在 $\alpha = 0.05$ 的显著性水平下，利用 F 检验去确定，在这些变量之间是否存在一个显著的关系。

24. 《华尔街日报》在全美一流大学中进行了一项有关篮球经费的研究。参加国家顶级篮球项目的39支球队的收入（100万美元）、获胜比例和教练薪酬（100万美元）的部分数据如下所示（*The Wall Street Journal*，2006年3月11～12日）

a. 建立了一个估计的回归方程，使该方程在球队参加国家篮球项目得到的收入和球队获胜的比例已知时，能用来预测球队教练的薪酬。

b. 在 $\alpha = 0.05$ 的显著性水平下，利用 F 检验来确定这些变量之间的关系在总体上的显著性。你的结论是什么？

c. 在 $\alpha = 0.05$ 的显著性水平下，利用 t 检验来确定每一个自变量的显著性。你的结论是什么？

学校名称	收入（100万美元）	获胜比例（%）	教练薪酬（100万美元）
Alabama	6.5	61	1.00
Arizona	16.6	63	0.70
Arkansas	11.1	72	0.80
Boston College	3.4	80	0.53
⋮	⋮	⋮	⋮
Washington	5.0	83	0.89
West Virginia	4.9	67	0.70
Wichita State	3.1	75	0.41
Wisconsin	12.0	66	0.70

26. 在第10题中，由美国职业棒球大联盟2011赛季的20名投手组成了一个随机样本，并且给出了这些投手的几个统计量数值。在第10题的（c）中，建立了一个估计的回归方程，该方程在投手每局比赛使对方击球员三击未中出局的平均次数和被对方击球员本垒打的平均次数已知时，能用来预测每局比赛的投手平均责任失分。

a. 在 $\alpha = 0.05$ 的显著性水平下，利用 F 检验来确定这些变量之间关系的总体显著性。你的结论是什么？

b. 在 $\alpha = 0.05$ 的显著性水平下，利用 t 检验来确定每一个自变量的显著性。你的结论是什么？

15.6 应用估计的回归方程进行估计和预测

在多元回归分析中，估计 y 的平均值和预测 y 的一个个别值的步骤，与包含一个自变量的回归分析所进行的那些步骤类似。首先让我们回忆一下，在第14章中我们已经说明了，对于 x 的一个给定值，y 的期望值的点估计和 y 的一个个别值的点估计是相同的。在两种情形下，我们都利用 $\hat{y} = b_0 + b_1x$ 作为它们的点估计。

在多元回归分析中，我们应用同样的步骤，即我们将自变量 x_1，x_2，…，x_p 的给定值代入估计的回归方程里，使用对应的 $\hat{y}$ 的值作为 y 的期望值和 y 的一个个别值的点估计。假定在 Butler 运输公司的例子里，我们希望利用含有 x_1（每天运送货物的行驶里程）和 x_2（运送货物的次数）两个自变量的估计的回归方程去建立两个区间估计：

1. 对于所有的运货汽车，在运送货物行驶100英里和运送货物2次的情形下，这些汽车平均行驶时间的置信区间。
2. 对于一辆特定的运货汽车，在运送货物行驶100英里和运送货物2次的情形下，该辆汽车行驶时间的预测区间。

利用估计的回归方程 $\hat{y} = -0.869 + 0.0611x_1 + 0.932x_2$，当 $x_1 = 100$ 和 $x_2 = 2$ 时，我们得到下面 $\hat{y}$ 的值。

$$\hat{y} = -0.869 + 0.0611 \times 100 + 0.932 \times 2 = 7.09$$

因此，在两种情形下，汽车行驶时间的点估计值大约等于7个小时。

为了建立y的平均值和y的一个个别值的区间估计，我们应用了与包含一个自变量的回归分析相类似的步骤。所需公式的推导已经超出了本教科书的范围，但是对于多元回归分析情形，一旦自变量x_1，x_2，…，x_p的值被用户设定，计算机软件包往往就能提供这些区间估计。在Butler运输公司的例子中，对于我们所选取的x_1和x_2的数值，表15-4给出了置信水平为95%的置信区间和预测区间的上、下限的数值；这些数值都可以通过使用Minitab得到。需要注意的是，y的一个个别值的区间估计要比y的期望值的区间估计有更宽的范围。这一区别只不过反映了这样一个事实：对于给定的x_1和x_2的数值，我们估计所有运货汽车的平均行驶时间比我们预测一辆特定运货汽车的行驶时间要有更高的精确度。

表15-4　Butler运输公司的置信水平为95%的置信区间和预测区间

x_1的数值	x_2的数值	置信区间		预测区间	
		下限	上限	下限	上限
50	2	3.146	4.924	2.414	5.656
50	3	4.127	5.789	3.368	6.548
50	4	4.815	6.948	4.157	7.607
100	2	6.258	7.926	5.500	8.683
100	3	7.385	8.645	6.520	9.510
100	4	8.135	9.742	7.362	10.515

练习

方法

☆28. 参阅在第2题中给出的数据。利用这些数据我们建立的估计的回归方程是

$$\hat{y} = -18.4 + 2.01x_1 + 4.74x_2$$

a. 当$x_1 = 45$，$x_2 = 15$时，建立一个y的平均值的置信水平为95%的置信区间。

b. 当$x_1 = 45$，$x_2 = 15$时，建立一个y的一个个别值的置信水平为95%的预测区间。

应用

30. 在第9题中，我们已经建立了一条小船的最高时速关于该船的最大宽度和发动机马力的估计的回归方程。

a. 对于宽度为85英寸和发动机马力为330的小船，建立一个平均最高时速的置信水平为95%的置信区间。

b. Svfara SV609的最大宽度是85英寸，发动机的马力是330。建立一个Svfara SV609最高时速的置信水平为95%的预测区间。

15.7　分类自变量

到目前为止，我们所考虑的例子的自变量都是定量自变量，例如学生人数、运送货物行驶的距离、运送货物的次数等。但是，在许多情形中，我们必须利用**分类自变量**（categorical independent variables）处理问题，例如性别（男、女），付款方式（现金、信用卡、支票），等等。我们这一节的目的是说明，如何在回归分析中应用分类变量。为了应用和解释分类自变量，我们将考虑约翰逊过滤股份公司（Johnson Filtration，Inc）的管理人员所面临的一个问题。[㊀]

15.7.1　一个例子：约翰逊过滤股份公司

约翰逊过滤股份公司对遍布南佛罗里达州的水过滤系统提供维修保养服务。当顾客的水过滤系统出现故障时，他们就要与约翰逊公司进行联系，请求公司对他们的水过滤系统进行维修。为了估计服务时间和服务成本，约翰逊公司的管理人员希望对顾客的每一次维修请求，预测必要的维修时间。所以，按小时计算的维修时间是应变量。管理人员认为，维修时间依赖两个因素：从最近一次维修服务至今水过滤系统已经使用的时间（单位：月）和需要维修的故障类型（机械的或电子的）。由10次维修服务组成一个样本，有关资料数据如表15-5所示。

我们用y表示按小时计算的维修时间，x_1表示从最近一次维修服务至今的时间，仅仅利用x_1进行预测的回归模型是

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

我们利用Minitab建立了估计的回归方程，得到的输出如图15-7所示。[㊁]估计的回归方程是

㊀ 自变量可以是分类变量，也可以是定量变量。

㊁ 在Minitab输出中，如果将变量名“Months”和“Time”作为列标题输入到工作表的单元格中，那么就有x_1 = Months和y = Time。

$$\hat{y} = 2.15 + 0.304x_1 \quad (15\text{-}16)$$

表 15-5 约翰逊过滤公司例子的数据

维修服务请求	最近一次维修服务至今的时间（月）	维修的故障类型	维修时间（小时）
1	2	电子	2.9
2	6	机械	3.0
3	8	电子	4.8
4	3	机械	1.8
5	2	电子	2.9
6	7	电子	4.9
7	9	机械	4.2
8	8	机械	4.8
9	4	电子	4.4
10	6	电子	4.5

回归方程是销售额
Time=2.15+0.304 Months

预测量	系数	标准差	t值	p-值
常数	2.147 3	0.605 0	3.55	0.008
Months	0.304 1	0.100 4	3.03	0.016

S=0.781 022 R–Sq=53.4% R–Sq(adj)=47.6%

方差分析

来源	自由度	平方和	均方	F	p
回归	1	5.596 0	5.596 0	9.17	0.016
误差	8	4.880 0	0.610 0		
总计	9	10.476 0			

图 15-7 在约翰逊过滤公司的例子中用最近一次维修服务至今的时间（x_1）作为自变量的 Minitab 输出

在 $\alpha=0.05$ 的显著性水平下，对于 t（或 F）检验，p-值为 0.016，这就表明了维修时间显著地依赖于从最近一次维修服务至今使用的时间。R – Sq =53.4% 表明了 x_1 仅仅解释了维修时间变异性的 53.4%。

为了将故障的维修类型引入回归模型，我们定义下面的变量。

$$x_2 = \begin{cases} 0, \text{如果故障的维修类型是机械的} \\ 1, \text{如果故障的维修类型是电子的} \end{cases}$$

在回归分析中，x_2 被称为**虚拟变量**（dummy variable）或指标变量。利用这个虚拟变量，我们能把多元回归模型写成如下形式

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

表 15-6 是经过修订后包含了虚拟变量数值的数据集。利用 Minitab 和表 15-6 的数据，我们能求出模型参数的估计值。图 15-8 所示的 Minitab 输出[⊖]表明了估计的多元回归方程是

$$\hat{y} = 0.93 + 0.388x_1 + 1.26x_2 \quad (15\text{-}17)$$

表 15-6 用虚拟变量表示维修故障类型（$x_2=0$ 表示机械的，$x_2=1$ 表示电子的）的约翰逊过滤公司例子的数据

维修服务请求	最近一次维修服务至今的时间（x_1）	维修的故障类型（x_2）	维修时间（y）
1	2	1	2.9
2	6	0	3.0
3	8	1	4.8
4	3	0	1.8
5	2	1	2.9
6	7	1	4.9
7	9	0	4.2
8	8	0	4.8
9	4	1	4.4
10	6	1	4.5

回归方程是销售额
Time=0.930+0.388 Months+1.26 Type

预测量	系数	标准差	t值	p-值
常数	0.930 5	0.467 0	1.99	0.087
Months	0.387 62	0.062 57	6.20	0.000
Type	1.262 7	0.314 1	4.02	0.005

S=0.459 048 R–Sq=85.9% R–Sq(adj)=81.9%

方差分析

来源	自由度	平方和	均方	F	p
回归	2	9.000 9	4.500 5	21.36	0.001
误差	7	1.475 1	0.210 7		
总计	9	10.476 0			

图 15-8 在约翰逊过滤公司的例子中用最近一次维修服务至今的时间（x_1）和维修类型（x_2）作为自变量的 Minitab 输出

在 $\alpha=0.05$ 的显著性水平下，与 F 检验（$F=21.36$）相联系的 p-值为 0.001，这就表明回归关系是显著的。在图 15-8 中，打印输出的 t 检验部分表明，从最近一次维修服务至今的时间（p-值 =0.000）和维修的类型（p-值 =0.005）这两个自变量在统计上都是显著的。另外，$R^2=85.9\%$ 和修正 $R^2=81.9\%$ 表明估计的回归方程很

⊖ 在 Minitab 输出中，如果将变量名"Months""Type"和"Time"作为列标题输入到工作表的单元格中，那么就有 $x_1=\text{Months}$、$x_2=\text{Type}$ 和 $y=\text{Time}$。

好地解释了维修时间的变异性。于是，对于各种不同类型的维修服务请求，在估计必要的维修时间时，估计的回归方程（15-17）应是有帮助的。

15.7.2 解释参数

对于约翰逊过滤公司的例子，多元回归方程是

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{15-18}$$

当模型中存在一个分类变量时，为了弄清楚如何对参数β_0，β_1和β_2作出合理的解释，我们考虑$x_2=0$（机械类型的故障）的情形。在已知故障类型是机械型时，用$E(y \mid \text{机械})$表示故障维修时间的平均值或期望值，我们有

$$E(y \mid \text{机械}) = \beta_0 + \beta_1 x_1 + \beta_2 \times 0 = \beta_0 + \beta_1 x_1 \tag{15-19}$$

类似地，对于电子类型的故障（$x_2=1$），我们有

$$E(y \mid \text{电子}) = \beta_0 + \beta_1 x_1 + \beta_2 \times 1 = \beta_0 + \beta_1 x_1 + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_1 \tag{15-20}$$

比较式（15-19）和（15-20），我们看到：无论是机械类型的故障还是电子类型的故障，平均维修时间都是x_1的线性函数。这两个方程的斜率都是β_1，但是y轴截距不同。对于机械类型的故障，在式（15-19）中的y轴截距是β_0；对于电子类型的故障，在式（15-20）中的y轴截距是（$\beta_0+\beta_2$）。β_2的解释是：它表示电子类型故障的平均维修时间与机械类型故障的平均维修时间之间的差。

如果β_2是正的，那么电子类型故障的平均维修时间将大于机械类型故障的平均维修时间；如果β_2是负的，那么电子类型故障的平均维修时间将小于机械类型故障的平均维修时间。最后，如果$\beta_2=0$，那么电子类型故障与机械类型故障的平均维修时间之间没有差别，维修时间不依赖于水过滤系统的故障类型。

利用估计的多元回归方程$\hat{y}=0.93+0.388x_1+1.26x_2$，我们看到：0.93是$\beta_0$的估计值，1.26是$\beta_2$的估计值。于是，当$x_2=0$（机械类型的故障）时，有

$$\hat{y} = 0.93 + 0.388x_1 \tag{15-21}$$

当$x_2=1$（电子类型的故障）时，有

$$\hat{y} = 0.93 + 0.388x_1 + 1.26 \times 1 = 2.19 + 0.388x_1 \tag{15-22}$$

实际上，对维修的故障类型引入虚拟变量为我们提供了能用于预测维修时间的两个估计的回归方程，一个方程对应机械类型故障的维修时间，一个方程对应电子类型故障的维修时间。另外，因为$b_2=1.26$，我们得知：电子类型故障的维修时间要比机械类型故障的维修时间平均多用了1.26个小时。

图15-9是根据表15-6的约翰逊过滤公司的数据画出的散点图。用纵轴表示按小时计算的维修时间（y），用水平轴表示从最近一次维修服务至今的月数（x_1）。机械类型故障的维修时间的数据点用M表示，电子类型故障的维修时间的数据点用E表示。在散点图上画出了式（15-21）和式（15-22）的图形，这就从几何上表明了这两个方程能用于预测维修时间，一个对应机械类型故障的维修时间，一个对应电子类型故障的维修时间。

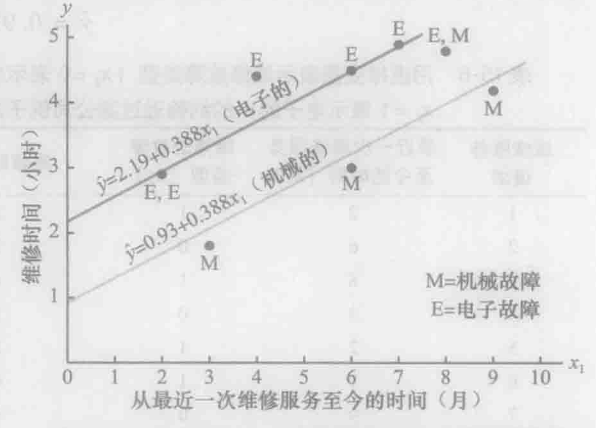


图15-9 根据表15-6的约翰逊过滤公司的维修数据作出的散点图

15.7.3 更复杂的分类变量

因为对于约翰逊过滤公司的例子，分类变量只有两个水平（机械的和电子的），我们只需要很容易地定义一个虚拟变量就可以了，用0表示需要维修的故障类型是机械型的，用1表示需要维修的故障类型是电子型的。但是，当一个分类变量超过两个水平时，我们必须在定义虚拟变量和解释虚拟变量两方面谨慎从事。正如我们将要说明的那样，如果一个分类变量有k个水平，那么需要定义$k-1$个虚拟变量，每一个虚拟变量或者取值为0，或者取值为1。[⊖]

⊖ 如果一个分类变量有k个水平，我们必须用$k-1$个虚拟变量来模拟这个分类变量。在定义和解释这些虚拟变量方面，我们必须谨慎行事。

例如，假定复印机制造商组织策划的销售区域是某一个特定州的A、B和C三个地区。管理人员希望，利用回归分析方法帮助他们去预测每周复印机的销售数量。用复印机的销售数量作应变量，并且考虑几个自变量（销售人员的数量，广告费用支出，等等）。我们还假定：管理人员相信，销售地区也是预测复印机销售数量的一个重要因素。因为，销售地区是一个分类变量，它有A、B和C三个水平，所以我们将需要3-1=2个虚拟变量来表示销售地区。每一个虚拟变量被定义为如下形式。

$$x_1 = \begin{cases} 1, \text{如果销售地区是B} \\ 0, \text{其他} \end{cases} \qquad x_2 = \begin{cases} 1, \text{如果销售地区是C} \\ 0, \text{其他} \end{cases}$$

根据这个定义，我们有x_1和x_2的值如下。

销售地区	x_1	x_2
A	0	0
B	1	0
C	0	1

对应于销售地区A的观测值，虚拟变量被定义为$x_1=0$，$x_2=0$；对应于销售地区B的观测值，虚拟变量被定义为$x_1=1$，$x_2=0$；对应于销售地区C的观测值，虚拟变量被定义为$x_1=0$，$x_2=1$。

复印机销售数量的期望值$E(y)$关于虚拟变量的回归方程将被写为

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

为了帮助我们理解参数β_0，β_1，和β_2的意义，我们考虑回归方程的三种变化。

$$E(y \mid \text{销售地区A}) = \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 = \beta_0$$

$$E(y \mid \text{销售地区B}) = \beta_0 + \beta_1 \times 1 + \beta_2 \times 0 = \beta_0 + \beta_1$$

$$E(y \mid \text{销售地区C}) = \beta_0 + \beta_1 \times 0 + \beta_2 \times 1 = \beta_0 + \beta_2$$

于是，β_0是地区A销售数量的平均值或期望值，β_1是地区B销售数量的平均值和地区A销售数量的平均值之间的差，β_2是地区C销售数量的平均值和地区A销售数量的平均值之间的差。

因为分类变量有3个水平，所以需要2个虚拟变量，但是，指定$x_1=0$，$x_2=0$表示销售地区A；指定$x_1=1$，$x_2=0$表示销售地区B；指定$x_1=0$，$x_2=1$表示销售地区C却是任意的。例如，我们也能用$x_1=1$，$x_2=0$表示销售地区A；用$x_1=0$，$x_2=0$表示销售地区B；用$x_1=0$，$x_2=1$表示销售地区C。在这种情形中，β_1能被解释为是地区A销售数量的平均值和地区B销售数量的平均值之间的差，β_2能被解释为是地区C销售数量的平均值和地区B销售数量的平均值之间的差。

重点是要牢记：在多元回归分析中，如果一个分类变量有k个水平，那么需要在多元回归分析中定义$k-1$个虚拟变量。于是，如果在销售地区的例子中还有第4个地区，我们不妨用D表示，那么有必要定义3个虚拟变量。例如，3个虚拟变量可以被定义为如下形式。

$$x_1 = \begin{cases} 1, \text{如果销售地区是B} \\ 0, \text{其他} \end{cases} \qquad x_2 = \begin{cases} 1, \text{如果销售地区是C} \\ 0, \text{其他} \end{cases} \qquad x_3 = \begin{cases} 1, \text{如果销售地区是D} \\ 0, \text{其他} \end{cases}$$

练习

方法

☆32. 考虑一个回归方程，它包含一个应变量y，一个定量自变量x_1和一个具有两个水平（水平1和水平2）的分类自变量。

a. 写出应变量y关于自变量x_1和分类变量的多元回归方程。

b. 对应于分类变量水平1，y的期望值是多少？

c. 对应于分类变量水平2，y的期望值是多少？

d. 对你的回归方程中的参数作出解释。

应用

☆34. 管理部门提出下面的回归方程用于预测一家快餐店的销售收入。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_2 + \varepsilon$$

式中，x_1代表在1英里范围内竞争者的个数；x_2代表在1英里范围内的居民人数（1 000人）；

$$x_3 = \begin{cases} 1, \text{如果为驾车者设置了不必下车的窗口} \\ 0, \text{否则} \end{cases}$$

y代表销售收入（1 000美元）。

在调查了20家快餐店之后，建立了下面的估计的回归方程。

$$\hat{y} = 10.1 - 4.2x_1 + 6.8x_2 + 15.3x_3$$

a. 由于为驾车者设置了不必下车就可以购买快餐食品的窗口而带来的期望销售收入是多少？

b. 若有一家快餐店，在1英里范围内有2个竞争者、8 000名居民并且没有为驾车者设置不必下车就可以购买快餐食品的窗口，预测这家快餐店的销售收入。

c. 若有一家快餐店，在1英里范围内有1个竞争者、3 000名居民并且为驾车者设置了不必下车就可以购买快餐食品的窗口，预测这家快餐店的销售收入。

36. 这个问题是第35题中所叙述情形的扩展。

a. 建立估计的回归方程，使得该方程在从最近一次维修服务至今的时间，维修的故障类型以及由哪一位维修工完成维修服务已知的条件下，能用来预测维修时间。

b. 在 $\alpha=0.05$ 的显著性水平下，对在（a）中所建立的估计的回归方程，检验在自变量和应变量之间是否存在一个显著的关系。

c. 增加维修服务是由哪一位维修工完成的自变量 x_3，这个自变量在统计上显著吗？检验的显著性水平为 $\alpha=0.05$。对于得到的结果，你能作出什么解释？

38. 美国心脏协会经过10年的研究，得到了有关中风风险与年龄、血压和吸烟嗜好相关性的统计资料。假设这一研究的部分数据如下表所示。我们将病人在今后十年内发生中风的概率（乘上100）视为中风的风险。我们用一个虚拟变量来定义病人是否为吸烟者，1表示是吸烟者，0表示是非吸烟者。

中风风险	年龄	血压	是否吸烟	中风风险	年龄	血压	是否吸烟
12	57	152	否	22	71	152	否
24	67	163	否	36	70	173	是
13	58	155	否	15	67	135	是
56	86	177	是	48	77	209	是
28	59	196	否	15	60	199	否
51	76	189	是	36	82	119	是
18	56	155	是	8	66	166	否
31	78	120	否	34	80	125	是
37	80	135	是	3	62	117	否
15	78	98	否	37	59	207	是

a. 建立一个中风风险关于人的年龄、血压和是否为吸烟者的估计的回归方程。

b. 在中风风险的估计的回归方程中，吸烟是一个显著的影响因素吗？检验的显著性水平为 $\alpha=0.05$。对于得到的结果，请作出解释。

c. Art Speen 是一位血压为175的68岁的吸烟者，他在今后十年内发生中风的概率是多少？对于这位病人，医生可以提出什么建议？

15.8 残差分析

在第14章中我们曾经指出，在作残差图和识别异常值时，经常要用到标准化残差。对于第 i 次观测，标准化残差的一般公式如下。

第 i 次观测的标准化残差

$$\frac{y_i - \hat{y}_i}{s_{y_i} - \hat{y}_i} \tag{15-23}$$

式中，$s_{y_i-\hat{y}_i}$ 表示第 i 次观测的残差的标准差。

第 i 次观测的残差的标准差的一般公式定义如下。

第 i 次观测的残差的标准差

$$s_{y_i-\hat{y}_i} = s\sqrt{1-h_i} \tag{15-24}$$

式中，s 代表估计的标准误差；h_i 代表第 i 次观测的杠杆率。

正如我们在第14章中已经陈述的那样，一次观测的**杠杆率**（leverage）是由自变量的观测值与这些观测值的

平均值之间的距离有多远确定的。在多元回归情形中，对于第 i 次观测，计算 h_i、$s_{y_i-\hat{y}_i}$ 和标准化残差，如果用手算是非常复杂的。但是，如果利用统计软件包，标准化残差可以很容易地从它的输出里得到。对于本章前面介绍的 Butler 运输公司的例子，表 15-7 列出了预测值、残差和标准化残差，我们通过利用 Minitab 统计软件包得到了这些数值。表中的预测值是根据估计的回归方程 $\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$ 计算出来的。

表 15-7　Butler 运输公司回归分析的残差和标准化残差

行驶距离（x_1）	送货次数（x_2）	行驶时间（y）	预测值（$\hat{y}$）	残差（$y-\hat{y}$）	标准化残差
100	4	9.3	8.938 46	0.361 541	0.783 44
50	3	4.8	4.958 30	−0.158 304	−0.349 62
100	4	8.9	8.938 46	−0.038 460	−0.083 34
100	2	6.5	7.091 61	−0.591 609	−1.309 29
50	2	4.2	4.034 88	0.165 121	0.381 67
80	2	6.2	5.868 92	0.331 083	0.654 31
75	3	7.4	6.486 67	0.913 331	1.689 17
65	4	6.0	6.798 75	−0.798 749	−1.773 72
90	3	7.6	7.403 69	0.196 311	0.367 03
90	2	6.1	6.480 26	−0.380 263	−0.776 39

图 15-10 是 Butler 运输公司多元回归例子的标准化残差图，在图中所使用的标准化残差和 y 的预测值 $\hat{y}$，都可以从表 15-7 中得到。这个标准化残差图并没有表示出任何与众不同的反常情形。同时，所有的标准化残差都介于 −2 和 +2 之间。因此，我们没有理由怀疑误差项 ε 是正态分布的假定。最后，我们的结论是：模型的假定都是合适的。

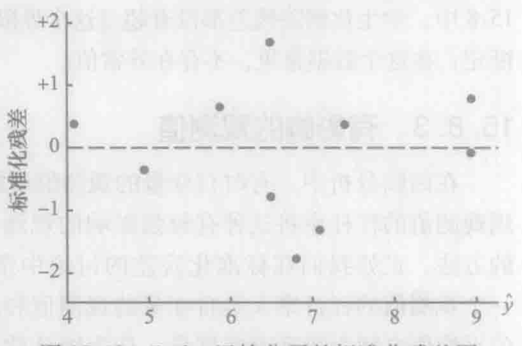


图 15-10　Butler 运输公司的标准化残差图

正态概率图也能用来确定误差项 ε 的分布是否服从正态分布的假定。正态概率图的作图步骤和解释已经在第 14.8 节中讨论过。对于多元回归分析，同样的步骤也是适用的。而且，我们仍然利用一个统计软件包来完成计算任务并绘制出正态概率图。

15.8.1　检测异常值

和其他数据相比较，**异常值**（outlier）是一个与众不同的观测值；换句话说，异常值与由其他观测值作出的散点图所显示的趋势不相吻合。在第 14 章中，我们已经给出了一个异常值的例子，并且讨论了如何利用标准化残差去检测异常值。如果一个观测值的标准化残差的数值小于 −2 或者大于 +2，Minitab 能将这个观测值识别为是一个异常值。对 Butler 运输公司例子的标准化残差（见表 15-7），应用这一规则，我们没有从数据集中检测出任何异常值。

一般来说，如果在数据集中存在一个或一个以上的异常值，将导致估计的标准误差 s 的增加，从而使得第 i 次观测的残差的标准差 $s_{y_i-\hat{y}_i}$ 的增加。因为在标准化残差公式（15-23）中的分母是 $s_{y_i-\hat{y}_i}$，所以标准化残差的数值将随着 s 值的增加而减少。最后导致的结果是，虽然一个残差可能超乎寻常得大，但由于在式（15-23）中有一个大的分母，从而可能使识别一个观测值为异常值的标准化残差规则失效。我们利用一种称为**学生化删除残差**（studentized deleted residual）的标准化残差，就能防止这种情形发生。

15.8.2　学生化删除残差和异常值

假设从数据集中删除第 i 次观测值，利用其余的 $n-1$ 次观测值建立一个新的估计的回归方程。设 $s_{(i)}$ 表示从数据集中删除了第 i 次观测值后得到的估计的标准误差。如果我们计算第 i 次观测的残差的标准差，用 $s_{(i)}$ 代替

式（15-24）中的 s，那么在我们计算第 i 次观测的标准化残差时，式（15-23）利用了 $s_{y_i-\hat{y}_i}$ 的修正值，这样得到的标准化残差称为学生化删除残差。如果第 i 次观测值是一个异常值，那么 $s_{(i)}$ 将小于 s。所以，第 i 次观测的学生化删除残差的绝对值将大于标准化残差的绝对值。在这种意义下，学生化删除残差可以检测出标准化残差不能检测出的异常值。

为了得到学生化删除残差，许多统计软件包提供了可选项。对于 Butler 运输公司例子，我们利用 Minitab 可以得到学生化删除残差，结果列在表 15-8 中。我们能利用 t 分布来确定，学生化删除残差是否表明存在异常值。回想一下，我们用 p 表示自变量的个数，用 n 表示观测值的个数。如果我们从数据集中删除第 i 次观测值，那么数据集中余下的观测值还有 $n-1$ 个；在这种情形下，误差平方和的自由度是 $(n-1)-p-1$。对于 Butler 运输公司例子，$n=10$，$p=2$，删除第 i 次观测值后，自由度是 $9-2-1=6$。在 0.05 显著性水平下，有 6 个自由度的 t 分布的双侧分位数 $t_{0.025}=2.447$（见附录 B 表 B-2）。如果第 i 次观测的学生化删除残差的数值小于 -2.447 或者大于 $+2.447$，我们就能断定：第 i 次观测值是一个异常值。因为在表 15-8 中，学生化删除残差都没有超过这个界限，所以我们断定：在这个数据集里，不存在异常值。

表 15-8　Butler 运输公司的学生化删除残差

行驶距离 (x_1)	送货次数 (x_2)	行驶时间 (y)	标准化残差	学生化删除残差
100	4	9.3	0.783 44	0.759 39
50	3	4.8	−0.349 62	−0.326 54
100	4	8.9	−0.083 34	−0.077 20
100	2	6.5	−1.309 29	−1.394 94
50	2	4.2	0.381 67	0.357 09
80	2	6.2	0.654 31	0.625 19
75	3	7.4	1.689 17	2.031 87
65	4	6.0	−1.773 72	−2.213 14
90	3	7.6	0.367 03	0.343 12
90	2	6.1	−0.776 39	−0.751 90

15.8.3　有影响的观测值

在回归分析中，有时自变量的观测值可能对回归结果有一个较强的影响，在第 14.9 节中，我们讨论了如何利用观测值的杠杆率将这种有较强影响的观测值识别出来的方法。正如我们在标准化残差的讨论中指出的那样，一个观测值的杠杆率 h_i 是自变量的观测值和这些观测值的平均值之间距离远近的度量。作为统计软件包输出的一部分，杠杆率的数值很容易就能得到。Minitab 不但能够计算出杠杆率的数值，而且还能够利用经验法则：$h_i>3(p+1)/n$ 识别出**有影响的观测值**（influential observations）。对于 Butler 运输公司的例子，自变量的个数是 $p=2$，观测值的个数是 $n=10$，杠杆率的临界值是 $3\times(2+1)/10=0.9$。利用 Minitab，我们得到的 Butler 运输公司例子的杠杆率数值如表 15-9 所示。因为 h_i 都不超过 0.9，所以在这个数据集里，我们没有检测出有影响的观测值。

表 15-9　Butler 运输公司的杠杆率和库克距离测度

行驶距离 (x_1)	送货次数 (x_2)	行驶时间 (y)	杠杆率 (h_i)	库克距离测度 (D_i)
100	4	9.3	0.351 704	0.110 994
50	3	4.8	0.375 863	0.024 536
100	4	8.9	0.351 704	0.001 256
100	2	6.5	0.378 451	0.347 923
50	2	4.2	0.430 220	0.036 663
80	2	6.2	0.220 557	0.040 381
75	3	7.4	0.110 009	0.117 562
65	4	6.0	0.382 657	0.650 029
90	3	7.6	0.129 098	0.006 656
90	2	6.1	0.269 737	0.074 217

15.8.4　利用库克距离测度识别有影响的观测值

利用杠杆率识别有影响的观测值时可能出现的一个问题是：一个观测值能被识别出来是一个有高杠杆率的观测值，但是根据得到的估计的回归方程，这个观测值未必是一个有影响的观测值。例如，表 15-10 是一个由 8 个观测值组成的数据集和它们对应的杠杆率（可以利用 Minitab 得到）。因为第 8 个观测值的杠杆率是 0.91 >

表 15-10　利用杠杆率的临界值说明潜在问题的数据集

x_i	y_i	h_i	x_i	y_i	h_i
1	18	0.204 170	4	23	0.125 977
1	21	0.204 170	4	24	0.125 977
2	22	0.164 205	5	26	0.127 715
3	21	0.138 141	15	39	0.909 644

0.75（杠杆率的临界值），所以这个观测值被识别出来是一个有影响的观测值。在得到最后的结论之前，让我们从不同的角度考虑这个问题。

图15-11给出了与表15-10的数据集相对应的散点图。我们利用Minitab建立了这些数据的估计的回归方程如下。

$$\hat{y} = 18.2 + 1.39x$$

图15-11中的直线就是这个方程的图示。现在让我们从数据集中将观测值（$x=15$，$y=39$）删除掉，对剩余的7个观测值拟合一个新的估计的回归方程。这个新的估计的回归方程是

$$\hat{y} = 18.1 + 1.42x$$

我们注意到，新的估计的回归方程和利用所有数据建立的估计的回归方程，这两个方程的y轴截距和斜率的数值没有显著的区别。尽管杠杆率准则将第8个观测值识别为是一个有影响的观测值，但显然这个观测值对我们的估计结果影响很小。于是在某些情形下，仅仅利用杠杆率来识别有影响的观测值，可能导致错误的结论。

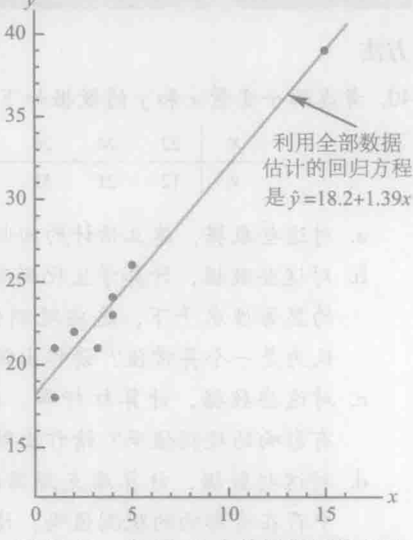


图15-11 表15-10中数据集的散点图

注：如果点（15，39）被删除掉，估计的回归方程是$\hat{y}=18.1+1.42x$

库克距离测度（Cook's distance measure）利用第i次观测的杠杆率h_i和第i次观测的残差（$y_i-\hat{y}_i$）来确定这个观测值是否是一个有影响的观测值。

库克距离测度

$$D_i = \frac{(y_i-\hat{y}_i)^2}{(p+1)\ s^2}\left[\frac{h_i}{(1-h_i)^2}\right] \tag{15-25}$$

式中，D_i代表第i次观测的库克距离测度；$y_i-\hat{y}_i$代表第i次观测的残差；h_i代表第i次观测的杠杆率；p代表自变量的个数；s代表估计的标准误差。

如果一个观测值的残差和（或）杠杆率比较大，那么库克距离测度也将比较大，从而就能表明这个观测值是一个有影响的观测值。作为经验准则，如果$D_i>1$，那么就表明第i次观测值是一个有影响的观测值，并应对这个观测值做进一步的考察。对于Butler运输公司的例子，表15-9的最后一列是由Minitab给出的库克距离测度。第8个观测值的库克距离测度$D_i=0.650\,029$，这个观测值应有最大的影响。但是根据$D_i>1$的经验准则，在Butler运输公司的数据集中，我们并不需要关注有影响观测值的存在问题。

注释和评论

1. 识别异常值和有影响的观测值的步骤预先向我们提出了警告，某些观测值可能对回归结果产生潜在的影响。我们对每一个异常值和有影响的观测值都要进行仔细检查。如果发现是数据的误差，那么应该对数据误差作出修正，并重新进行回归分析。一般地说，我们不应该将异常值和有影响的观测值从数据集中删除掉，除非有明显的证据表明，它们不是我们正在研究的总体元素，这时候我们就不应该再把它们包括在原始数据集中。
2. 为了确定库克距离测度D_i的值是否足够大，使得我们能得出第i次观测值是一个有影响的观测值的结论，我们还能将D_i的值与分子自由度为$p+1$、分母自由度为$n-p-1$的F分布的第50个百分位数（表示为$F_{0.50}$）做比较。对应于显著性水平$\alpha=0.50$的F分布表一定能实现这一检验。我们给出的经验准则（$D_i>1$）就是根据以下事实：对于相当广泛的各种不同情形，F分布表的值非常接近于1。

练习

方法

40. 考虑两个变量 x 和 y 的数据如下。

x_i	22	24	26	28	40
y_i	12	21	31	35	70

a. 对这些数据，建立估计的回归方程。

b. 对这些数据，计算学生化删除残差。在 $\alpha = 0.05$ 的显著性水平下，这些观测值中的哪一个能被认为是一个异常值？请作出解释。

c. 对这些数据，计算杠杆率。在这些数据中存在有影响的观测值吗？请作出解释。

d. 对这些数据，计算库克距离测度。在这些数据中存在有影响的观测值吗？请作出解释。

应用

42. 下面是16款流行的Sports/GT型车的自重、马力和从开始启动并加速至1/4英里处时的速度的数据。假设每一款Sports/GT型车的价格也能得到。全部数据集如下表所示。

a. 建立一个估计的回归方程，使该方程在价格和马力已知时，能用来预测从开始启动并加速至1/4英里处时的速度。

b. 绘制关于$\hat{y}$的标准化残差图。残差图是否证实了有关误差项 ε 的假定？请作出解释。

c. 对这些数据，检查异常值。你的结论是什么？

d. 是否存在有影响的观测值？请作出解释。

高性能跑轿车	价格（1 000 美元）	自重（磅）	马力	在1/4英里处时的速度（英里/小时）
Acura Integra Type R	25.035	2 577	195	90.7
Acura NSX-T	93.758	3 066	290	108.0
BMW Z3 2.8	40.900	2 844	189	93.2
ChevroletCamaro Z28	24.865	3 439	305	103.2
Chevrolet Corvette Convertible	50.144	3 246	345	102.1
Dodge Viper RT/10	69.742	3 319	450	116.2
Ford Mustang GT	23.200	3 227	225	91.7
Honda Prelude Type SH	26.382	3 042	195	89.7
Mercedes-Benz CLK320	44.988	3 240	215	93.0
Mercedes-Benz SLK230	42.762	3 025	185	92.3
Mitsubishi 3000GT VR-4	47.518	3 737	320	99.0
Nissan 240SX SE	25.066	2 862	155	84.6
Pontiac Firebird Trans Am	27.770	3 455	305	103.2
PorscheBoxster	45.560	2 822	201	93.2
Toyota Supra Turbo	40.989	3 505	320	105.0
Volvo C70	41.120	3 285	236	97.0

15.9 logistic 回归

在许多回归应用中，应变量仅能够被假定为是两个离散值。例如，为了预测人们是否愿意使用某种信用卡，一家银行很想建立一个估计的回归方程。如果银行批准使用某种信用卡的申请，则应变量被赋值为 $y=1$；如果银行拒绝使用某种信用卡的申请，则应变量被赋值为 $y=0$。利用 logistic 回归，在被选取出来的自变量的一组特定值已知的条件下，我们能估计银行将批准使用某种信用卡申请的概率。

现在让我们考虑 logistic 回归的一个应用，该应用涉及了 Simmons 商店正在使用的一种直接邮寄广告的促销手段。Simmons 拥有并且经营一家全美国范围的妇女服饰连锁商店。这家连锁店印制了 5 000 份成本昂贵的彩色商品目录，并且每份商品目录还赠送一张商家的优惠券，如果顾客购买 200 美元或 200 美元以上的商品，将给予 50 美元的折扣。因为商品目录的成本昂贵，所以 Simmons 只愿意将商品目录寄送给那些最有可能使用优惠券并购买 200 美元或 200 美元以上商品的顾客。

管理人员认为，在预测收到商品目录的顾客是否将使用优惠券并购买 200 美元或 200 美元以上的商品时，两个可能会有所帮助的变量是在 Simmons 商店的年消费支出和顾客是否拥有 Simmons 信用卡。Simmons 进行了一项试验性的研究，随机样本由 50 名拥有 Simmons 信用卡的顾客和 50 名没有 Simmons 信用卡的其他顾客组成。对被随机抽取出来

的100名顾客，Simmons把商品目录送给他们中的每一名顾客。在一个试验周期结束时，Simmons记录下顾客是否使用了优惠券。前10名商品目录接受者的样本数据如表15-11所示。每一名顾客去年在Simmons商店的年消费支出以1 000美元为单位；信用卡的信息被赋值，如果顾客拥有Simmons信用卡，则被赋值为1；否则为0。优惠券的信息也被赋值，如果样本中的顾客使用了优惠券并购买了200美元或200美元以上的商品，则被赋值为1；否则为0。

表15-11 Simmons商店例子的部分样本数据

顾客	年消费支出（1 000美元）	Simmons信用卡	优惠券	顾客	年消费支出（1 000美元）	Simmons信用卡	优惠券
1	2.291	1	0	6	2.473	0	1
2	3.215	1	0	7	2.384	0	0
3	2.135	1	0	8	7.076	0	0
4	3.924	0	0	9	1.182	1	1
5	2.528	1	0	10	3.345	0	0

我们可以利用表15-11的数据建立一个多元回归模型，该模型能帮助Simmons预测：一名商品目录的接受者是否将使用优惠券并购买200美元或200美元以上的商品。我们用年消费支出和Simmons信用卡作为自变量，优惠券作为应变量。因为应变量仅可以被假定为两个值，0或1，所以无法应用常规的多元回归模型。这个例子说明，在怎样的情形下，应建立logistic回归这种类型的模型。下面我们将看到，如何利用logistic回归去帮助Simmons预测哪一种类型的顾客最有可能利用商店直接邮寄广告的促销手段。

15.9.1 logistic回归方程

在许多方面，logistic回归与普通回归是类似的。logistic回归需要一个应变量y和一个或一个以上的自变量。在多元回归分析中，我们把y的平均值或期望值称为多元回归方程。

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{15-26}$$

在logistic回归中，我们不但在实践上，而且在统计理论上也已经证明了：下面的非线性方程较好地描述了$E(y)$和x_1，x_2，…，x_p之间的关系。

logistic回归方程

$$E(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}} \tag{15-27}$$

如果应变量y的两个值被赋值为0或1，那么在自变量x_1，x_2，…，x_p的一组特定值已知的条件下，式（15-27）中$E(y)$的值给出了$y=1$的概率。由于$E(y)$被解释为概率，所以通常将**logistic回归方程**（logistic regression equation）写成如下形式。

logistic回归中的$E(y)$被解释为概率

$$E(y) = P(y = 1 \mid x_1, x_2, \cdots, x_p) \tag{15-28}$$

为了更好地理解logistic回归方程的特征，假定模型仅包括一个自变量x，并且模型的参数值为$\beta_0 = -7$，$\beta_1 = 3$。对应这些参数值，logistic回归方程为

$$E(y) = P(y = 1 \mid x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{e^{-7+3x}}{1 + e^{-7+3x}} \tag{15-29}$$

图15-12是式（15-29）的图示。注意，图形为一条S形曲线。$E(y)$的取值范围为0～1，当自变量x的值变得愈来愈大时，则$E(y)$的值逐渐地增加，并且愈来愈接近于1；当自变量x的值变得愈来愈小时，则$E(y)$的值逐渐地减少，并且愈来愈

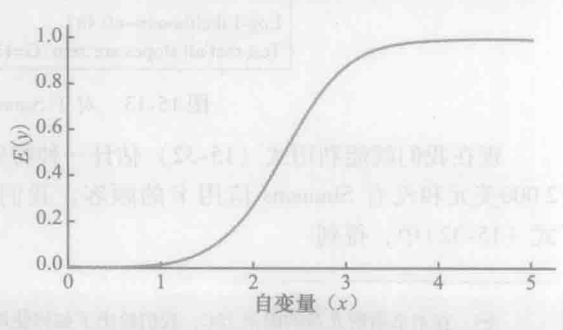


图15-12 对于$\beta_0 = -7$，$\beta_1 = 3$的logistic回归方程

接近于0。还需要注意的是，当自变量x的值从2增加到3时，表示概率的$E(y)$的值增加得非常快。由于$E(y)$的取值范围为0~1和$E(y)$为一条S形曲线这两个事实，就从理论上使式（15-29）非常适合做应变量的概率等于1的模型。

15.9.2 估计 logistic 回归方程

在简单线性回归和多元回归时，我们利用最小二乘法计算模型参数（β_0，β_1，β_2，…，β_p）的估计量b_0，b_1，b_2，…，b_p。因为logistic回归方程具有非线性形式，因此计算估计量的方法要更复杂，并且超出了本教科书的范围。我们将利用计算机软件包求出这些估计量，得到了下面的**估计的logistic回归方程**（estimated logistic regression equation）。

估计的 logistic 回归方程

$$\hat{y}=P\ (y=1\mid x_1,\ x_2,\ \cdots,\ x_p)\ 的估计=\frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1+e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}} \tag{15-30}$$

式中，$\hat{y}$是在自变量x_1，x_2，…，x_p的一组特定值已知时，给出了$y=1$的概率的估计。

现在让我们回到Simmons商店的例子。在这个问题中，我们定义的变量如下所示。

$$y=\begin{cases}0, 如果顾客没有使用优惠券\\1, 如果顾客使用了优惠券\end{cases}$$

$$x_1=在\ \text{Simmons}\ 商店的年消费支出(1\,000\ 美元)$$

$$x_2=\begin{cases}0, 如果顾客没有\ \text{Simmons}\ 信用卡\\1, 如果顾客有\ \text{Simmons}\ 信用卡\end{cases}$$

于是，我们选择有两个自变量的logistic回归方程。

$$E(y)=\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2}} \tag{15-31}$$

我们利用样本数据（见表15-11）和Minitab的二进制logistic回归程序计算模型参数β_0，β_1，β_2的估计值。[⊖]得到的Minitab的部分输出如图15-13所示。[⊜]我们看到，$b_0=-2.146\,37$，$b_1=0.341\,643$和$b_2=1.098\,73$。于是估计的logistic回归方程是

$$\hat{y}=\frac{e^{b_0+b_1x_1+b_2x_2}}{1+e^{b_0+b_1x_1+b_2x_2}}=\frac{e^{-2.146\,37+0.341\,643x_1+1.098\,73x_2}}{1+e^{-2.146\,37+0.341\,643x_1+1.098\,73x_2}} \tag{15-32}$$

Logistic Regression Table

Predictor	Coef	SE Coef	Z	p	Odds Ratio	95% Lower	CI Upper
Constant	−2.146 37	0.577 245	−3.72	0.000			
Spending	0.341 643	0.128 672	2.66	0.008	1.41	1.09	1.81
Card	1.098 73	0.444 696	2.47	0.013	3.00	1.25	7.17

Log-Likelihood=−60.487
Test that all slopes are zero: G=13.628, DF=2, P-Value=0.001

图15-13 对于Simmons商店例子的logistic回归的部分输出

现在我们就能利用式（15-32）估计一种特定类型的顾客使用优惠券的概率。例如，对于去年的消费支出为2 000美元和没有Simmons信用卡的顾客，我们估计他们使用优惠券的概率。为此我们将$x_1=2$，$x_2=0$代入式（15-32)中，得到

⊖ 在本书所附光盘的附录15C，我们给出了如何使用Minitab生成如图15-13所示的输出。
⊜ 在Minitab输出中，x_1 = Spending 和 x_2 = Card。

$$\hat{y} = \frac{e^{-2.146\,37+0.341\,643\times 2+1.098\,73\times 0}}{1+e^{-2.146\,37+0.341\,643\times 2+1.098\,73\times 0}} = \frac{e^{-1.643\,1}}{1+e^{-1.463\,1}} = \frac{0.231\,5}{1.231\,5} = 0.188\,0$$

于是，对于这种特定类型的顾客群体，他们使用优惠券的概率大约为0.19。类似地，我们可以估计去年的消费支出为2 000美元和拥有Simmons信用卡的顾客，他们使用优惠券的概率。为此，我们将 $x_1=2$，$x_2=1$ 代入式（15-32），得到

$$\hat{y} = \frac{e^{-2.146\,37+0.341\,643\times 2+1.098\,73\times 1}}{1+e^{-2.146\,37+0.341\,643\times 2+1.098\,73\times 1}} = \frac{e^{-0.364\,4}}{1+e^{-0.364\,4}} = \frac{0.694\,6}{1.694\,6} = 0.409\,9$$

于是，对于这种特定类型的顾客群体，他们使用优惠券的概率大约为0.41。这一结果显示出：拥有Simmons信用卡的顾客，他们使用优惠券的概率要比没有Simmons信用卡的顾客大得多。但是，在推断出任何的结论之前，我们必须对得到的估计的logistic回归方程的统计显著性进行评估。

15.9.3 显著性检验

logistic回归的显著性检验与多元回归的显著性检验相类似。首先，我们进行总体的显著性检验。对于Simmons商店的例子，总体的显著性检验如下：

$$H_0: \beta_1 = \beta_2 = 0$$

H_a：至少有一个参数不等于零

总体的显著性检验是根据 G 检验统计量的值进行的。如果原假设成立，则 G 统计量的抽样分布为服从自由度等于模型中自变量个数的 χ^2 分布。虽然 G 统计量的计算超出了本教科书的范围，但是 G 统计量的值和它所对应的 p-值，已在Minitab的二元logistic回归的输出部分给出。我们从图15-13的最后一行看到 G 统计量的值是13.628，它的自由度是2，它的 p-值是0.001。于是，在 $\alpha \geq 0.001$ 的任一显著性水平下，我们应该拒绝原假设，并且得到模型的总体是显著的结论。

如果 G 检验表明模型的总体是显著的，那么我们就可以利用 z 检验来确定每一个单个的自变量对模型总体是否有显著的作用。对于自变量 x_i，提出的假设是

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

如果原假设成立，则估计的系数 b_i 除以它的标准差 s_{b_i} 后，得到的结果 $z_i = b_i / s_{b_i}$ 为一服从标准正态分布的统计量。在Minitab输出中的列标题"Z"包含了每一个被估计的系数的 z_i 值，列标题"p"包含了对应的 p-值。假定我们在 $\alpha=0.05$ 的显著性水平下，检验Simmons商店模型中自变量的显著性。对于自变量 x_1，z 值是2.66，对应的 p-值是0.008。于是，在 $\alpha=0.05$ 的显著性水平下，我们应该拒绝 H_0：$\beta_1=0$。用类似的方法，我们也应该拒绝 H_0：$\beta_2=0$，这是因为对应于 $z=2.47$ 的 p-值是0.013。所以在 $\alpha=0.05$ 的显著性水平下，两个自变量在统计上都是显著的。

15.9.4 管理上的应用

我们已经讨论了如何建立估计的logistic回归方程和如何检验它的显著性。现在让我们利用这些结果对Simmons商店通过邮寄商品目录进行促销的问题作出一个决策建议。对于Simmons商店，我们已经计算出了 $P(y=1 \mid x_1=2, x_2=1)=0.409\,9$ 和 $P(y=1 \mid x_1=2, x_2=0)=0.188\,0$。这些概率表明：对于去年的消费支出为2 000美元的顾客，拥有Simmons信用卡增加了使用优惠券购买商品的概率。对去年的消费支出1 000～7 000美元的两类顾客：一类拥有Simmons信用卡，另一类没有Simmons信用卡，我们在表15-12中看到了他们使用优惠券的估计的概率。Simmons将如何利用这些信息，为顾客制定新的促销策略？假定Simmons只愿意将促销的商品目录邮寄给那些使用优惠券的概率等于或大于0.40的顾客。利用表15-12中估计的概率，Simmons促销策略应该是：

拥有Simmons信用卡的顾客：对去年的消费支出为2 000美元或2 000美元以上的每一位顾客都邮寄促销的商品目录。

没有 Simmons 信用卡的顾客：对去年的消费支出为6 000 美元或6 000 美元以上的每一位顾客都邮寄促销的商品目录。

表 15-12 Simmons 商店的估计的概率

	去年的消费支出（美元）						
	1 000	2 000	3 000	4 000	5 000	6 000	7 000
拥有信用卡	0.330 5	0.409 9	0.494 3	0.579 1	0.659 4	0.731 5	0.793 1
没有信用卡	0.141 3	0.188 0	0.245 7	0.314 4	0.392 2	0.475 9	0.561 0

我们再进一步留意一下估计的概率，我们看到，对于没有 Simmons 信用卡，但是去年的消费支出为5 000 美元的顾客，他们使用优惠券的概率是 0.392 2。于是，Simmons 或许觉得需要修订这一促销策略，即将那些虽然没有 Simmons 信用卡，只要去年的消费支出为5 000 美元或5 000 美元以上的顾客也包括进来。

15.9.5 解释 logistic 回归方程

对一个回归方程作出解释涉及与我们建立方程要回答的商务问题有关的自变量。利用 logistic 回归，直接地解释自变量和 $y=1$ 的概率之间的关系非常困难，这是因为 logistic 回归方程是非线性方程。但是统计学家们已经证明：他们能利用一个被称为机会比率的概念间接地解释这一关系。

有利于一个事件发生的机会比（odds in favor of an event occurring）被定义为事件将要发生的概率与该事件将不会发生的概率的比。在 logistic 回归中，我们总是将感兴趣的事件设定为 $y=1$。于是在自变量的一组特定值已知时，有利于事件 $y=1$ 发生的机会比能按下式计算：

$$\text{机会比} = \frac{P(y=1 \mid x_1, x_2, \cdots, x_p)}{P(y=0 \mid x_1, x_2, \cdots, x_p)} = \frac{P(y=1 \mid x_1, x_2, \cdots, x_p)}{1-P(y=1 \mid x_1, x_2, \cdots, x_p)} \tag{15-33}$$

机会比率（odds ratio）度量了当一组自变量中只有一个自变量增加了一个单位时，对机会比的影响。机会比率是当给定的一组自变量中的一个自变量增加了一个单位时，$y=1$ 的机会比（odds_1）除以该组自变量的值都没有变化时，$y=1$ 的机会比（odds_0）。

机会比率

$$\text{机会比率} = \frac{\text{odds}_1}{\text{odds}_0} \tag{15-34}$$

例如，假定我们想要对去年的消费支出为2 000 美元并且拥有 Simmons 信用卡的顾客（$x_1=2$，$x_2=1$）使用优惠券的机会比，与去年的消费支出为2 000 美元但没有 Simmons 信用卡的顾客（$x_1=2$，$x_2=0$）使用优惠券的机会比进行比较。我们感兴趣的是：解释在自变量 x_2增加一个单位时的影响。在这种情形时，

$$\text{odds}_1 = \frac{P(y=1 \mid x_1=2, x_2=1)}{1-P(y=1 \mid x_1=2, x_2=1)}$$

和

$$\text{odds}_0 = \frac{P(y=1 \mid x_1=2, x_2=0)}{1-P(y=1 \mid x_1=2, x_2=0)}$$

我们在前面已经得到，在 $x_1=2$，$x_2=1$ 时，$y=1$ 的概率的估计值是 0.409 9，$x_1=2$，$x_2=0$ 时，$y=1$ 的概率的估计值是 0.188 0。于是，

$$\text{odds}_1 \text{ 的估计值} = \frac{0.409\,9}{1-0.409\,9} = 0.694\,6$$

$$\text{odds}_0 \text{ 的估计值} = \frac{0.188\,0}{1-0.188\,0} = 0.231\,5$$

估计的机会比率是

$$估计的机会比率 = \frac{0.6946}{0.2315} = 3.00$$

于是，我们能够断定：去年的消费支出为2 000美元并且拥有Simmons信用卡的顾客使用优惠券的机会比是去年的消费支出同为2 000美元但没有Simmons信用卡的顾客使用优惠券的机会比的3倍。

当所有其他的自变量都保持不变时，我们能计算出每一个自变量的机会比率。但是所有其他的自变量取什么样的常数值却是无关紧要的。例如，如果用3 000美元代替去年2 000美元的消费支出（x_1），我们对Simmons信用卡变量（x_2）计算机会比率，我们还会得到相同的机会比率的估计值（3.00）。于是，我们能够断定：拥有Simmons信用卡的顾客使用优惠券的机会比是没有Simmons信用卡的顾客使用优惠券的机会比的3倍。

对于logistic回归软件包，机会比率是标准输出。让我们参照图15-13 Minitab的部分输出进行说明。列标题"Odds Ratio"包含了每一个自变量的估计的机会比率。x_1的估计的机会比率是1.41，x_2的估计的机会比率是3.00。我们已经说明了应该如何解释二进制的自变量x_2的估计的机会比率。现在让我们考虑如何解释连续型自变量x_1的估计的机会比率。

在Minitab输出的列标题"Odds Ratio"下，数值1.41告诉我们：去年消费支出为3 000美元的顾客使用优惠券的估计的机会比是去年消费支出为2 000美元的顾客使用优惠券的估计的机会比的1.41倍。而且，这个解释对x_1的任何一个单位的变化都成立。例如，去年消费支出为5 000美元的顾客使用优惠券的估计的机会比是去年消费支出为4 000美元的顾客使用优惠券的估计的机会比的1.41倍。但是，假定我们感兴趣的问题是，当自变量的增加超过一个单位时，机会比的变化情况。注意，x_1的取值范围是1~7。作为Minitab的输出被打印出来的机会比率不能回答这个问题。为了回答这个问题，我们必须考察机会比率和回归系数之间的关系。

在一个变量的机会比率和它所对应的回归系数之间存在一个唯一的关系。在一个logistic回归方程中，每一个自变量都能表示成如下形式

$$机会比率 = e^{\beta_i}$$

为了说明这个关系，考虑在Simmons商店例子中的自变量x_1。x_1的估计的机会比率是

$$估计的机会比率 = e^{b_1} = e^{0.341643} = 1.41$$

类似地，x_2的估计的机会比率是

$$估计的机会比率 = e^{b_2} = e^{1.09873} = 3.00$$

一旦我们求出了模型参数的估计值，利用自变量的机会比率和回归系数之间的关系，就能很容易地计算出机会比率的估计值。而且它也为我们提供了当连续型自变量的变化大于或小于一个单位时，研究机会比率变化的能力。

当一个自变量变化一个单位，而所有其他的自变量都保持不变时，一个自变量的机会比率描述了该自变量机会比的变化。假定，我们想要考虑自变量的变化大于一个单位，比如c个单位时的影响。例如，假定在Simmons商店例子中，我们想要对去年的消费支出为5 000美元（$x_1=5$）的顾客使用优惠券的机会比，与去年的消费支出为2 000美元（$x_1=2$）的顾客使用优惠券的机会比进行比较。在这种情形时，$c=5-2=3$，对应的估计的机会比率是

$$e^{cb_1} = e^{3\times 0.341643} = e^{1.0249} = 2.79$$

这个结果表明：去年消费支出为5 000美元的顾客使用优惠券的估计的机会比是去年消费支出为2 000美元的顾客使用优惠券的估计的机会比的2.79倍。换句话说，对于一个去年消费支出增加3 000美元的顾客而言，使用优惠券的估计的机会比率是2.79。

一般说来，机会比率使我们能够比较两个不同事件的机会比。如果机会比率的值是1，那么两个事件的机会比是相同的。于是，如果正在考虑的自变量（例如Simmons信用卡的状态）对事件发生的概率有一个正影响，那么对应的机会比率将大于1。大多数的logistic回归软件包都给出了机会比率的置信区间。图15-13的Minitab输出对每一个机会比率都给出了置信水平为95%的置信区间。例如x_1的机会比率的点估计值是1.41，置信水平为95%的置信区间是（1.09，1.81）。因为置信区间不包含数值1，所以我们能够断定：x_1对估计的机会比率有一个显著的影响。类似地，x_2的机会比率的置信水平为95%的置信区间是（1.25，7.17）。因为这个区间也不包含数值1，所以我们也能够断定：x_2对估计的机会比率有一个显著的影响。

15.9.6 对数机会比（logit）变换

在有利于 $y=1$ 的机会比和 logistic 回归方程中 e 的指数之间，我们能够观察到一个有趣的关系。这个关系可以表示为

$$\ln(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

这个方程表明：有利于 $y=1$ 的机会比的自然对数是自变量的线性函数。我们称这个线性函数为**对数机会比**（logit）。我们用符号 $g(x_1, x_2, \cdots, x_p)$ 表示对数机会比。

对数机会比

$$g(x_1, x_2, \cdots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{15-35}$$

用 $g(x_1, x_2, \cdots, x_p)$ 代替式（15-27）中的 $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$，我们能将 logistic 回归方程写成如下形式

$$E(y) = \frac{e^{g(x_1, x_2, \cdots, x_p)}}{1 + e^{g(x_1, x_2, \cdots, x_p)}} \tag{15-36}$$

一旦我们估计出 logistic 回归方程的参数，我们就能计算出对数机会比的估计值。我们用符号 $\hat{g}(x_1, x_2, \cdots, x_p)$ 表示**估计的对数机会比**（estimated logit），从而我们就得到

估计的对数机会比

$$\hat{g}(x_1, x_2, \cdots, x_p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \tag{15-37}$$

于是，根据估计的对数机会比，估计的 logistic 回归方程是

$$\hat{y} = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}} = \frac{e^{\hat{g}(x_1, x_2, \cdots, x_p)}}{1 + e^{\hat{g}(x_1, x_2, \cdots, x_p)}} \tag{15-38}$$

对于 Simmons 商店的例子，估计的对数机会比是

$$\hat{g}(x_1, x_2) = -2.146\,37 + 0.341\,643 x_1 + 1.098\,73 x_2$$

估计的 logistic 回归方程是

$$\hat{y} = \frac{e^{\hat{g}(x_1, x_2)}}{1 + e^{\hat{g}(x_1, x_2)}} = \frac{e^{-2.146\,37 + 0.341\,643 x_1 + 1.098\,73 x_2}}{1 + e^{-2.146\,37 + 0.341\,643 x_1 + 1.098\,73 x_2}}$$

于是，由于在估计的对数机会比和估计的 logistic 回归方程之间存在唯一的关系，所以我们能计算 Simmons 商店的概率，这个概率可以通过用 $e^{\hat{g}(x_1, x_2)}$ 除以 $1 + e^{\hat{g}(x_1, x_2)}$ 得到。

注释和评论

1. 由于在模型中被估计的系数和对应的机会比率之间存在唯一的关系，建立在 G 统计量基础上的总体的显著性检验也是机会比率的总体显著性检验。另外，对模型单个参数显著性的 z 检验也给出了对应的机会比率显著性的统计检验。
2. 在简单和多元回归情形中，我们利用判定系数来度量模型的拟合优度。在 logistic 回归情形中，没有单一的测度能给出类似的解释。拟合优度的讨论已经超出了我们关于 logistic 回归介绍性论述的范围。

练习

应用

44. 参阅在本节中介绍的 Simmons 商店的例子。如果顾客使用优惠券购买了商品，则应变量被赋值为 $y=1$，否则，$y=0$。假定，唯一能帮助我们预测顾客是否将使用优惠券可得到的有用信息是顾客拥有信用卡的状况，如果顾客拥有 Simmons 信用卡，则自变量被赋值为 $x=1$，否则，$x=0$。

    a. 写出 y 关于 x 的 logistic 回归方程。

b. 当 $x=0$ 时，如何解释 $E(y)$？

c. 对于表15-11中Simmons商店的数据，利用Minitab计算估计的对数机会比（estimated logit）。

d. 利用在（c）中得到的估计的对数机会比，计算没有Simmons信用卡的顾客使用优惠券购买商品的概率的估计值，以及计算拥有Simmons信用卡的顾客使用优惠券购买商品的概率的估计值。

e. 机会比率的估计值是多少？它的解释是什么？

46. Community Bank想增加使用工资单直接存款的客户人数。管理部门考虑一个新的销售计划，该计划将要求每一个部门的管理人员电话联系每一位目前没有使用工资单直接存款的客户。作为登记报名参加使用工资单直接存款计划的一种奖励，将为每一位有联系的客户提供两年的免费账目核对。由于新计划涉及时间和费用方面的原因，管理部门想把他们的努力集中在登记报名参加使用工资单直接存款计划概率最高的那些客户身上。管理部门相信：客户支票账户上的平均每月的余额可能是该客户是否参加使用工资单直接存款计划的有帮助的预测器。为了查明这两个变量之间的关系，Community Bank抽取了50位目前没有使用工资单直接存款的支票账户客户组成了一个随机样本，然后利用这个样本来试验新的计划。样本数据给出了支票账户平均每月的余额（1 000美元）和管理人员电话联系过的客户是否登记报名参加使用工资单直接存款计划（如果客户登记报名参加使用工资单直接存款计划，则被赋值1，否则被赋值0）。这些数据被存放在本书所附光盘名为Bank的数据集里；部分数据如下表所示。

客户	x = 支票账户平均每月余额（1 000美元）	y = 直接存款
1	1.22	0
2	1.56	0
3	2.10	0
4	2.25	0
5	2.89	0
6	3.55	0
7	3.56	0
8	3.65	1
⋮	⋮	⋮
48	18.45	1
49	24.98	0
50	26.05	1

a. 写出 y 关于 x 的logistic回归方程。

b. 对于Community Bank的数据，利用Minitab计算估计的logistic回归方程。

c. 在 $\alpha=0.05$ 的显著性水平下，利用 G 检验统计量，进行显著性检验。

d. 如果客户平均每月支票账户的余额为1 000美元，估计该客户将登记报名参加使用工资单直接存款计划的概率。

e. 假定Community Bank只希望与登记报名参加使用工资单直接存款计划的概率大于或等于0.5的客户进行电话联系。要达到这个要求的概率水平，客户的平均每月支票账户的余额应是多少？

f. 机会比率的估计值是多少？它的解释是什么？

48. *Tire Rack* 杂志保持了一项独立的消费者调查工作，通过司机们的互相帮助，来分享他们长期使用轮胎的经验。对68条全季节轮胎的调查结果存放在本书所附光盘名为TireRatings的文件中（*Tire Rack* website，2012年3月21日）。对轮胎的性能特点使用下面的10分制评定等级：

优		良		好		一般		差	
10	9	8	7	6	5	4	3	2	1

标记为Wet的变量的数值是每条轮胎湿牵引性能的平均等级，标记为Noise的变量的数值是每条轮胎产生的噪声水平的平均等级。受访者还被问道，使用下面的10分制评定等级，他们是否会再次购买轮胎：

肯定		大概		可能		大概不会		绝对不	
10	9	8	7	6	5	4	3	2	1

标记为Buy Again的变量的数值是回答再次购买的平均值。为了这道习题的目的，我们建立下面的二进制应变量：

$$\text{Purchase}=\begin{cases}1, \text{如果变量 Buy Again 的值} \geq 7\\0, \text{如果变量 Buy Again 的值} < 7\end{cases}$$

于是，如果Purchase = 1，受访者大概或肯定会再次购买轮胎。

a. 如果 x_1 表示变量Wet的性能等级，x_2 表示变量Noise的性能等级，y 表示变量Purchase，写出 y 关于 x_1 和 x_2 的logistic回归方程。

b. 利用Minitab计算估计的对数机会比。

c. 利用估计的对数机会比，对于变量Wet和Noise的性能等级都是8的一条特定的轮胎，计算一名顾客

大概或肯定会再次购买该轮胎的概率的估计值。

d. 假设变量 Wet 和 Noise 的性能等级都是 7。这对一名顾客大概或肯定会再次购买具有这些性能等级的一条特定轮胎的概率会有什么影响？

e. 如果你是一家轮胎公司的 CEO，(c) 和 (d) 中的结果告诉你什么？

小结

作为在第 14 章中介绍的简单线性回归分析的推广，在这一章里，我们引入了多元回归分析的内容。多元回归分析能够使我们了解一个应变量是如何依赖两个或两个以上自变量的。回归方程 $E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$ 描述了应变量 y 的平均值或期望值是如何依赖自变量 x_1，x_2，…，x_p 的值。我们利用样本数据和最小二乘法建立了估计的回归方程 $\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p$。实际上，$b_0$，$b_1$，$b_2$，…，$b_p$ 是用来估计模型的未知参数 β_0，β_1，β_2，…，β_p 的样本统计量。遍及全章，我们都用到了计算机打印输出，这是为了强调这样一个事实：统计软件包是完成多元回归分析所需要的烦琐计算的唯一实用工具。

作为估计的回归方程拟合优度的一个度量，我们介绍了多元判定系数的概念。多元判定系数测定了在应变量 y 的变异性中，能被估计的回归方程解释的部分所占的比例。修正多元判定系数是一个类似的拟合优度的度量，它用自变量的个数进行调整，这样就可以避免由于增加较多自变量而产生的高估影响。

在这一章里我们还介绍了 F 检验和 t 检验，这两个检验都是判断变量间的关系在统计上是否显著的工具。F 检验用于确定应变量和所有自变量之间在总体上是否存在一个显著的关系。t 检验用于确定，在回归模型中的其他自变量已知时，应变量和单个的自变量之间是否存在一个显著的关系。自变量之间的相关性被称为多重共线性，我们已经讨论过这一概念。

分类自变量这一节叙述了如何利用虚拟变量将分类数据与多元回归分析结合起来。残差分析这一节叙述了如何利用残差分析来证实模型的假设、检测异常值和识别有影响的观测值的方法。在这一节里，我们还讨论了标准化残差、杠杆率、学生化删除残差和库克距离测度等问题。作为本章的结束，我们用一节的篇幅介绍了 logistic 回归问题，在这一节里，我们讨论了如何利用 logistic 回归来模拟应变量可能仅有两个假定值的情形。

关键术语

multiple regression analysis 多元回归分析 包含两个或两个以上自变量的回归分析。

multiple regression model 多元回归模型 描述应变量 y 是如何依赖自变量 x_1，x_2，…，x_p 和误差项 ε 的数学方程。

multiple regression equation 多元回归方程 应变量的期望值或平均值与自变量的值之间关系的数学方程，即 $E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$。

estimated multiple regression equation 估计的多元回归方程 根据样本数据和最小二乘法得到的多元回归方程的估计，即 $\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p$。

least squares method 最小二乘法 用来建立估计的回归方程的方法。这个方法使残差（应变量的观测值 y_i 和应变量的估计值 $\hat{y}$ 之间的离差）的平方和极小化。

multiple coefficient of determination 多元判定系数 估计的多元回归方程拟合优度的度量。它被理解为在应变量 y 的变异性中，能被估计的回归方程解释的部分所占的比例。

adjusted multiple coefficient of determination 修正多元判定系数 估计的多元回归方程拟合优度的度量，它用自变量的个数进行调整，这样就可以避免由于增加较多自变量而产生的高估影响。

multicollinearity 多重共线性 用于描述自变量之间相关性的术语。

categoricalindependent variable 分类自变量 使用分类数据的自变量。

dummy variable 虚拟变量 用于模拟分类自变量影响的变量。虚拟变量的取值可以仅为 0 或 1。

leverage 杠杆率 自变量的值与自变量的平均值之间距离远近的度量。

outlier 异常值 与其他数据所显示的趋势不相吻合的观测值。

studentized deleted residuals 学生化删除残差 通过从数据集中删除第 i 次观测值，然后完成回归分析

和有关计算，得到一个修正的估计的标准误差，在修正的估计的标准误差的基础上得到的标准化残差。

influential observation 有影响的观测值 对回归结果有一个强影响的观测值。

Cook's distance measure 库克距离测度 建立在第 i 次观测的杠杆率和第 i 次观测残差的基础上的观测影响程度的度量。

logistic regression equation logistic 回归方程 关于 $E(y)$ 的数学方程，而 $E(y)$ 是对于自变量 $x_1, x_2, \cdots, x_p$ 的值，$y=1$ 的概率，即 $E(y)=P(y=1 \mid x_1, x_2, \cdots, x_p)=\dfrac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}$。

estimated logistic regression equation 估计的 logistic 回归方程 根据样本数据得到的 logistic 回归方程的估计，即 $\hat{y}=P(y=1 \mid x_1, x_2, \cdots, x_p)$ 的估计 $=\dfrac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1+e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}$。

odds in favor of an event occurring 有利于一个事件发生的机会比 事件将要发生的概率与该事件将不会发生的概率的比。

odds ratio 机会比率 当一组自变量中只有一个自变量增加了一个单位时，$y=1$ 的机会比（$odds_1$）除以该组自变量的值没有变化时，$y=1$ 的机会比（$odds_0$），即机会比率 $=odds_1/odds_0$。

logit 对数机会比 有利于事件 $y=1$ 发生的机会比的自然对数，即 $g(x_1, x_2, \cdots, x_p)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$。

estimated logit 估计的对数机会比 根据样本数据得到的对数机会比的估计，即 $\hat{g}(x_1, x_2, \cdots, x_p)=b_0+b_1x_1+b_2x_2+\cdots+b_px_p$。

重要公式

多元回归模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon \quad (15\text{-}1)$$

多元回归方程

$$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p \quad (15\text{-}2)$$

估计的多元回归方程

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p \quad (15\text{-}3)$$

最小二乘法准则

$$\min\sum(y_i-\hat{y}_i)^2 \quad (15\text{-}4)$$

SST，SSR 和 SSE 之间的关系

$$SST=SSR+SSE \quad (15\text{-}7)$$

多元判定系数

$$R^2=\frac{SSR}{SST} \quad (15\text{-}8)$$

修正多元判定系数

$$R_a^2=1-(1-R^2)\frac{n-1}{n-p-1} \quad (15\text{-}9)$$

均方回归

$$MSR=\frac{SSR}{p} \quad (15\text{-}12)$$

均方误差

$$MSE=\frac{SSE}{n-p-1} \quad (15\text{-}13)$$

F 检验统计量

$$F=\frac{MSR}{MSE} \quad (15\text{-}14)$$

t 检验统计量

$$t=\frac{b_i}{s_{b_i}} \quad (15\text{-}15)$$

第 i 次观测的标准化残差

$$\frac{y_i-\hat{y}_i}{s_{y_i-\hat{y}_i}} \quad (15\text{-}23)$$

第 i 次观测的残差的标准差

$$s_{y_i-\hat{y}_i}=s\sqrt{1-h_i} \quad (15\text{-}24)$$

库克距离测度

$$D_i=\frac{(y_i-\hat{y}_i)^2}{(p+1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right] \quad (15\text{-}25)$$

logistic 回归方程

$$E(y)=\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}} \quad (15\text{-}27)$$

估计的 logistic 回归方程

$\hat{y}=P(y=1 \mid x_1,x_2,\cdots,x_p)$ 的估计

$$=\frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1+e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}} \quad (15\text{-}30)$$

机会比率

$$\text{机会比率}=\frac{odds_1}{odds_0} \quad (15\text{-}34)$$

对数机会比

$$g(x_1,x_2,\cdots,x_p)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p \quad (15\text{-}35)$$

估计的对数机会比

$$\hat{g}(x_1,x_2,\cdots,x_p)=b_0+b_1x_1+b_2x_2+\cdots+b_px_p \quad (15\text{-}37)$$

补充练习

50. 电子联营公司的人事主管建立了雇员对工作满意程度的测试成绩与他或她的工龄和工资率之间关系的估计的回归方程如下。

$$\hat{y} = 14.4 - 8.69x_1 + 13.5x_2$$

式中，x_1为工龄（年）；x_2为工资率（美元）；y为工作满意程度的测试成绩（比较高的分数表示对工作的满意程度也比较高）。

a. 解释在这个估计的回归方程中的系数。

b. 对于有4年工龄，每小时的工资为6.5美元的一名雇员，预测他对工作满意程度的测试成绩。

52. 参阅第49题，Clearwater大学招生办公室工作人员建立了该校学生大学毕业考试的GPA关于SAT的数学成绩和高中阶段的GPA的估计的回归方程如下。

$$\hat{y} = -1.41 + 0.0235x_1 + 0.00486x_2$$

式中，x_1代表高中阶段的GPA；x_2代表SAT的数学成绩；y代表大学毕业考试的GPA。

Minitab的部分计算机输出如下。

回归方程是$\hat{y} = -1.41 + 0.0235x_1 + 0.00486x_2$

预测量	系数	系数的标准差	t值
常数	-1.4053	0.4848	
x_1	0.023467	0.008666	____
x_2	____	0.001077	

$s = 0.1298$ R^2 = ____% R_a^2 = ____%

方差分析

来源	自由度	平方和	均方	F
回归	____	1.76209	____	____
误差	____	____	____	
总的	9	1.88000		

a. 计算Minitab输出中空缺的数字。

b. 利用F检验，在$\alpha = 0.05$的显著性水平下，检验变量间是否存在一个显著的关系。

c. 利用t检验，在$\alpha = 0.05$的显著性水平下，检验H_0: $\beta_1 = 0$和H_0: $\beta_2 = 0$。

d. 估计的回归方程对观测数据的拟合好吗？请作出解释。

54. 在美国，Tire Rack是处于领先地位的轮胎和车轮在线经销商，为了向客户提供令人满意的产品，该经销商进行了全面的测试，使这些产品均能满足客户对车辆、驾驶风格和驾驶条件的要求。此外，Tire Rack还保持了一项独立的消费者调查工作，通过司机们的互相帮助，来分享他们长期使用轮胎的经验。下面是对18条最高性能的夏季轮胎的调查等级分（等级分为1~10，1表示最差，10表示最好）。变量“转向”评定的是轮胎的转向反应性等级，变量“胎面磨损”是根据司机的预期评定的轮胎胎面磨损速率的等级，变量“再次购买”评定的是司机对轮胎的整体满意度，以及希望再次购买同样轮胎愿望的等级（Tire Rack website，2009年2月3日）。

轮胎	转向	胎面磨损	再次购买
Goodyear Assurance TripleTred	8.9	8.5	8.1
MichelinHydroEdge	8.9	9.0	8.3
Michelin Harmony	8.3	8.8	8.2
Dunlop SP 60	8.2	8.5	7.9
Goodyear AssuranceComfor Tred	7.9	7.7	7.1
Yokohama Y372	8.4	8.2	8.9
Yokohama Aegis LS4	7.9	7.0	7.1
Kumho Power Star 758	7.9	7.9	8.3
Goodyear Assurance	7.6	5.8	4.5
Hankook H406	7.8	6.8	6.2
Michelin Energy LX4	7.4	5.7	4.8
Michelin MX4	7.0	6.5	5.3
Michelin Symmetry	6.9	5.7	4.2
Kumho 722	7.2	6.6	5.0
Dunlop SP 40 A/S	6.2	4.2	3.4
Bridgestone Insignia SE200	5.7	5.5	3.6
Goodyear Integrity	5.7	5.4	2.9
Dunlop SP20 FE	5.7	5.0	3.3

a. 建立一个估计的回归方程，使该方程在轮胎转向反应性的等级已知时，能用来预测再次购买同样轮胎愿望的等级。在$\alpha = 0.05$的显著性水平下，检验变量间是否存在一个显著的关系。

b. 在（a）中建立的估计的回归方程对观测数据的拟合好吗？请作出解释。

c. 建立一个估计的回归方程，使该方程在轮胎转向反应性的等级和轮胎胎面磨损速率的等级已知时，能用来预测再次购买同样轮胎愿望的等级。

d. 在$\alpha = 0.05$的显著性水平下，检验增加的“胎面磨损”自变量是显著的自变量吗？

56. 下面是包含45只共同基金有关信息的数据集的一部分，这些共同基金是2008年晨星基金500（*Morningstar Funds* 500）的一部分。完整的数据集可以从本书所附光盘名为MutualFunds的文件中得到。数据集包括以下5个变量：

基金类型：国内股票、国际股票和固定收益。

资产净值（美元）：2007年12月31日的每股收盘价。

5年平均收益率（%）：过去5年，基金的平均年收率益。

费用比率（%）：每个会计年度从资产中扣除的基金费用的比例。

晨星排名：每一只基金的风险调整星级；晨星排名从最低的1星到最高的5星。

基金名称	基金类型	资产净值（美元）	5年平均收益率（%）	费用比率（%）	晨星排名
Amer Cent Inc & Growth Inv	国内股票	28.88	12.39	0.67	2星
American Century Intl. Disc	国际股票	14.37	30.53	1.41	3星
American Century Tax-Free Bond	固定收益	10.73	3.34	0.49	4星
American Century Ultra	国内股票	24.94	10.88	0.99	3星
Ariel	国内股票	46.39	11.32	1.03	2星
Artisan Intl Val	国际股票	25.52	24.95	1.23	3星
Artisan Small Cap	国内股票	16.92	15.67	1.18	3星
Baron Asset	国内股票	50.67	16.77	1.31	5星
Brandywine	国内股票	36.58	18.14	1.08	4星
⋮	⋮	⋮	⋮	⋮	⋮

a. 建立一个估计的回归方程，使该方程在基金类型已知时，能用来预测过去5年基金的平均收益率。在$\alpha=0.05$的显著性水平下，检验变量间关系的显著性。

b. 在（a）中建立的估计的回归方程对观测数据的拟合好吗？请作出解释。

c. 建立一个估计的回归方程，使该方程在基金类型、资产净值和费用比率已知时，能用来预测过去5年基金的平均收益率。在$\alpha=0.05$的显著性水平下，检验变量间关系的显著性。你觉得，是否有变量应从估计的回归方程中删除？请作出解释。

d. 晨星排名是一个分类变量。因为数据中只包含四个星级（从2星到5星），所以利用下面的虚拟变量：3StarRank = 1，如果是一只3星基金，否则，3StarRank = 0；4StarRank = 1，如果是一只4星基金，否则，4StarRank = 0；5StarRank = 1，如果是一只5星基金，否则，5StarRank = 0。建立一个估计的回归方程，使该方程在基金类型、费用比率和晨星排名已知时，能用来预测过去5年的基金平均收益率。在$\alpha=0.05$的显著性水平下，删除任何不显著的自变量。

e. 对于一只费用比率为1.05%，晨星排名为3星级的国内股票基金，利用在（d）中建立的估计的回归方程，预测该基金过去5年的平均收益率。

案例15-1　消费者调查股份有限公司

消费者调查股份有限公司是一家独立的机构，该机构为各种类型的厂商调查消费者的态度和行为。在一项研究中，公司的客户要求对消费者的特点进行调查，其目的是为了能预测消费者使用信用卡进行支付的数额。对于由50名消费者组成的一个样本，采集了有关年收入、家庭成员人数和年信用卡支付数额的统计资料。下表所示的数据存放在本书所附光盘名为Consumer的文件中。

年收入（1 000美元）	家庭成员人数	信用卡支付数额（美元）	年收入（1 000美元）	家庭成员人数	信用卡支付数额（美元）	年收入（1 000美元）	家庭成员人数	信用卡支付数额（美元）
54	3	4 016	21	2	2 448	62	3	4 705
30	2	3 159	44	1	2 995	64	2	4 157
32	4	5 100	37	5	4 171	22	3	3 579
50	5	4 742	62	6	5 678	29	4	3 890
31	2	1 864	21	3	3 623	39	2	2 972
55	2	4 070	55	7	5 301	35	1	3 121
37	1	2 731	42	2	3 020	39	4	4 183
40	2	3 348	41	7	4 828	54	3	3 730
66	4	4 764	54	6	5 573	23	6	4 127
51	3	4 110	30	1	2 583	27	2	2 921
25	3	4 208	48	2	3 866	26	7	4 603
48	4	4 219	34	5	3 586	61	2	4 273
27	1	2 477	67	4	5 037	30	2	3 067
33	2	2 514	50	2	3 605	22	4	3 074
65	3	4 214	67	5	5 345	46	5	4 820
63	4	4 965	55	6	5 370	66	4	5 149
42	6	4 412	52	2	3 890			

管理报告

1. 利用描述统计学的方法整理这些数据。对你的发现进行评述。

2. 首先用年收入作为自变量，然后用家庭成员人数作为自变量，分别建立估计的回归方程。那一个自变量是更好地预测年信用卡支付数额的变量？讨论你的发现。

3. 用年收入和家庭成员人数作为自变量，建立估计的回归方程。讨论你的发现。

4. 对于年收入为40 000美元的三口之家，该家庭预测的年信用卡支付数额是多少？

5. 请对模型是否需要增加其他的自变量进行讨论。增加自变量可能会有什么帮助？

案例15-2 预测NASCAR车手的奖金

MattKenseth赢得了纳斯卡赛车（NASCAR，全美运动汽车竞赛协会的缩写）2012年赛季最重要的比赛Daytona 500。他的获胜并不令人意外，因为他在2011赛季的积分榜上，以2 330分名列第四。落后于Tony Stewart（2 403分），Carl Edwards（2 403分）和Kevin Harvick（2 345分）。在2011年，Matt Kenseth挣了6 183 580美元的奖金，因为他得三次首发（排位赛最快的车手），获得三场比赛胜利，12次进入前五名，20次进入前十名。在2011年，NASCAR的积分系统规定，第一名车手得到43个积分，第二名车手得到42个积分，等等，直到第四十三名车手得到1个积分。另外，领先一圈的任何车手得到1个奖励积分，领先最多圈的车手得到额外的奖励积分，并且给比赛获胜者奖励3个积分。但是，一名车手在任何比赛中，最多能够得到48个积分。表15-13是2011赛季前35名车手的数据（NASCAR website，2011年2月28日）。

表15-13 2011赛季NASCAR成绩

车手	积分	首发	获胜	前5	前10	奖金(美元)	车手	积分	首发	获胜	前5	前10	奖金(美元)
Tony Stewart	2 403	1	5	9	19	6 529 870	Marcos Ambrose	936	0	1	5	12	4 750 390
Carl Edwards	2 403	3	1	19	26	8 485 990	Jeff Burton	935	0	0	2	5	3 807 780
Kevin Harvick	2 345	0	4	9	19	6 197 140	Juan Montoya	932	2	0	2	8	5 020 780
Matt Kenseth	2 330	3	3	12	20	6 183 580	Mark Martin	930	2	0	2	10	3 830 910
Brad Keselowski	2 319	1	3	10	14	5 087 740	David Ragan	906	2	1	4	8	4 203 660
Jimmie Johnson	2 304	0	2	14	21	6 296 360	Joey Logano	902	2	0	4	6	3 856 010
Dale Earnhardt Jr.	2 290	1	0	4	12	4 163 690	Brian Vickers	846	0	0	3	7	4 301 880
Jeff Gordon	2 287	1	3	13	18	5 912 830	Ragan Smith	820	0	1	2	5	4 579 860
Denny Hamlin	2 284	0	1	5	14	5 401 190	Jamie McMurray	795	1	0	2	4	4 794 770
Ryan Newman	2 284	3	1	9	17	5 303 020	David Reutimann	757	1	0	1	3	4 374 770
Kurt Busch	2 262	3	2	8	16	5 936 470	Boddy Labonte	670	0	0	1	2	4 505 650
Kyle Busch	2 246	1	4	14	18	6 161 020	David Gilliland	572	0	0	1	2	3 878 390
Clint Bowyer	1 047	0	1	4	16	5 633 950	Casey Mears	541	0	0	0	0	2 838 320
Kasey Kahne	1 041	2	1	8	15	4 775 160	Dave Blaney	508	0	0	1	1	3 229 210
A. J. Allmendinger	1 013	0	0	1	10	4 825 560	Andy Lally	398	0	0	0	0	2 868 220
Greg Biffle	997	3	0	3	10	4 318 050	Robby Gordon	268	0	0	0	0	2 271 890
Paul Menard	947	0	1	4	8	3 853 690	J. J. Yeley	192	0	0	0	0	2 559 500
Martin Truex Jr.	937	1	0	3	12	3 955 560							

管理报告

1. 假设你想使用车手获得首发的次数、获胜的次数、进入前五名的次数和进入前十名的次数这四个变量中的一个来预测该车手挣得的奖金（美元）。这四个变量中的哪一个能给出最佳的预测值？

2. 建立一个估计的回归方程，使该方程在车手获得首发的次数、获胜的次数、进入前五名的次数和进入前十名的次数已知时，能用来预测该车手挣得的奖金（美元）。对每一个自变量做单个的显著性检验。讨论你的发现和结论。

3. 建立两个新的自变量：Top2-5 和 Top6-10(Top2-5 表示车手进入第2位至第5位的次数，Top6-10 表示车手进入第6位至第10位的次数)。建立一个估计的回归方程，使该方程在车手获得首发的次数、获胜的次数、Top2-5 和 Top6-10 已知时，能用来预测该车手挣得的奖金（美元）。对每一个自变量做单个的显著性检验。讨论你的发现和结论。

4. 根据你的分析结果，为了预测车手的奖金（美元），你推荐哪个估计的回归方程？对该方程的估计的回归系数作出解释。

案例 15-3 找到最合适的汽车价值

当你试图决定购买一辆汽车时，决定汽车实际价值的并不一定是你在初次购买时花了多少钱。相反，性能可靠的并且不用花很多钱就能买到的汽车，往往代表了最合适的价值。但是，不管汽车的性能多么可靠，或者购买汽车的价钱多么便宜，它还必须要有很好的操控能力。

为了度量汽车的价值，《消费者报告》杂志提出了一个被称为价值分数的统计量。价值分数是根据汽车车主五年的费用、汽车道路综合测试分数以及预测可靠性等级得出的。车主五年的费用是该车在第一个五年所产生的各项支出，包括折旧，燃油，维修及保养，等等。根据一辆车每年行驶12 000 英里的全国平均水平，得到每英里的平均费用作为车主五年费用的测度。道路测试分数是超过50 次测试和评估的结果，并且以100 分为满分来计算成绩，较高的分数表示有较好的性能、舒适性、便利性及燃油经济性。由《消费者报告》杂志所实施的道路测试，得到的最高分是一辆雷克萨斯 LS 460L，为 99 分。预测可靠性等级（1 = 差，2 = 一般，3 = 好，4 = 很好，5 = 优秀）是基于 *Consumer Reports*' Annual Auto Survey 的数据得到的。

价值分数为1.0 的汽车被认为是“平均值”。价值分数为2.0 的汽车被认为是比价值分数为1.0 的汽车好两倍的汽车；价值分数为0.5 的汽车被认为只有价值分数为1.0 的汽车的一半好；等等。在本书所附光盘名为 CarValuede 的文件中有三种规格轿车（13 辆小型轿车、20 辆家庭轿车和21 辆高档轿车）的数据，包括每辆被测试轿车的价格（美元）（*Consumer Reports* website，2012 年4 月18 日）。由于加入了汽车规格的影响，考虑一个三水平（小型轿车、家庭轿车和高档轿车）的分类变量，我们需要利用两个虚拟变量如下：

$$\text{家庭轿车} = \begin{cases} 1,\text{如果是家庭轿车} \\ 0,\text{否则} \end{cases}$$

$$\text{高档轿车} = \begin{cases} 1,\text{如果是高档轿车} \\ 0,\text{否则} \end{cases}$$

管理报告

1. 以“费用/英里”为应变量，以“家庭轿车”和“中高档轿车”为自变量，建立一个估计的回归方程。讨论你的发现。

2. 以“价值分数”为应变量，以“费用/英里”“道路测试分”“预测可靠性等级”“家庭轿车”和“中高档轿车”为自变量，建立一个估计的回归方程。

3. 在 $\alpha = 0.05$ 的显著性水平下，从在（2）建立的估计的回归方程中，删除不显著的自变量。删除不显著的自变量后，建立一个新的估计的回归方程。

4. 假设有人主张：“最小的汽车比较大的汽车提供了更好的价值”。对于本例中的数据，变量“Small Sedans”代表最小的汽车，变量“Upscale Sedans”代表较大型的汽车。你的分析支持这种说法吗？

5. 利用回归分析建立一个估计的回归方程，使该方程在道路测试分已知时，能用来预测该车的价值分数。

6. 利用回归分析建立一个估计的回归方程，使该方程在预测的可靠性等级已知时，能用来预测该车的价值分数。

7. 从你的分析中，你可以得出什么结论？

第 16 章
回归分析：建立模型

CHAPTER 16

实践中的统计

Monsanto 公司[⊖]

密苏里州，圣路易斯

Monsanto 公司始建于 1901 年，当时约翰 F. 奎尼（John F. Queeney）依靠 500 美元起家，在密西西比河边一座布满尘土的破旧仓库，开始生产糖精。如今，该公司已经成为美国最大的化学公司之一，生产 1 000 多种各类产品，从工业化学制品到用于铺设现代运动场所需要的人工合成地面，应有尽有。Monsanto 公司也是一家世界性的跨国公司，在 65 个国家拥有生产设备、实验室、技术中心和营销机构。

Monsanto 公司的营养化学部生产和销售一种蛋氨酸补充剂，这一产品广泛用于肉鸡、猪、牛等家畜的饲料。由于肉鸡养殖业产量大且利润率低，因此，需要营养价值尽可能高的有成本效益的肉鸡饲料。最优的饲料成分将可以满足，在一个给定的喂食量上，使得幼鸡兼具快速的生长率和最终的高标准体重。Monsanto 公司的化学工业部门与饲养者们紧密合作，以生产出最优的肉鸡饲料。该公司最终的成功取决于他们发现了肉鸡饲料的最优配方，使得肉鸡的饲养成本与猪、牛等其他家畜的饲养成本相比，保持在更低的水平上。

Monsanto 公司运用回归分析方法模拟了肉鸡的体重 y 与饲料中加入的蛋氨酸数量 x 之间的关系。最初建立的估计的简单线性回归方程如下：

$$\hat{y} = 0.21 + 0.42x$$

这一估计的回归方程经检验在统计上是显著的，但是残差分析显示，曲线关系也许是描述肉鸡体重与饲料中加入的蛋氨酸数量之间关系的一个更适宜的模型。

Monsanto 公司经过进一步的研究发现，虽然少量的蛋氨酸可以提高肉鸡的体重，但是当其含量达到某一水平后，继续增加蛋氨酸的含量对肉鸡体重增长的作用就变得微乎其微了。事实上，当蛋氨酸含量的增加超出了营养需求量时，肉鸡的体重甚至会下降。下面估计的多元回归方程用于模拟肉鸡体重与饲料中加入的蛋氨酸数量之间的曲线关系。

$$\hat{y} = -1.89 + 1.32x - 0.506x^2$$

利用这一回归结果，Monsanto 公司能够确定在肉鸡饲料中需要添加蛋氨酸的最优数量。

在本章中，我们将通过对诸如 Monsanto 公司建模的例子，扩展回归分析的讨论至非线性模型。此外，我们还要介绍一系列工具，这些工具有助于决定最佳回归方程中应该包含哪些自变量。

建立模型是一个复杂的过程，经过这一过程，我们就能得到描述应变量与一个或一个以上自变量之间关系的估计的回归方程。建模过程主要有两方面的问题，一是找到一个合适的描述变量之间关系的函数形式，二是选择模型应包含的自变量。在第 16.1 节中，我们将通过引入一般线性模型的概念，确立建模过程的基本框架。在第 16.2 节中，我们将介绍什么时候增加或者删除自变量的一般方法，为比较复杂的基本计算步骤提供了依据。在第 16.3 节中，我们将考虑一个涉及 8 个自变量和 25 组观测值的大型回归问题。在第 16.4 节中，这个大型问题将用于说明变量选择过程，包括逐步回归过程、前向选择过程、后向消元过程和最佳子集回归过程等。在第 16.5 节中，我们将说明多元回归分析可以提供处理实验设计问题的另一种方法。在第 16.6 节中，我们将说明如何利用杜宾 - 瓦特森（Durbin-Watson）检验来检测序列相关或者自相关。

16.1　一般线性模型

假设我们采集了一个应变量 y 和 k 个自变量 x_1，x_2，…，x_k 的观测数据。我们的目的是利用这些数据，建立一个估计的回归方程，这个方程给出了应变量和自变量之间一个最佳的关系。作为建立自变量之间更复杂关系的总体框架，我们介绍含有 p 个自变量的**一般线性模型**（general linear model）的概念。

⊖ 作者感谢 Monsanto 公司营养化学部的两位高级研究专家 James R. Ryland 和 Robert M. Schisla 先生，他们为“实践中的统计”提供了本案例。

一般线性模型

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \varepsilon \tag{16-1}$$

在式（16-1）中，每一个自变量 z_j（这里，$j=1, 2, \cdots, p$）都是 $x_1, x_2, \cdots, x_k$（我们已经采集了这些变量的数据）的函数。[⊖]在某种情形下，每一个自变量 z_j 可能仅仅是一个变量 x 的函数。最简单的情形是我们仅仅对一个变量 x_1 采集数据，并且希望利用直线关系去估计 y。在这种情形下，$z_1 = x_1$，式（16-1）变为

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \tag{16-2}$$

式（16-2）恰好是第 14 章中介绍的简单线性回归模型，不同之处是将第 14 章模型的自变量 x 换成 x_1。在统计建模文献中，这个模型被称为具有一个预测变量的简单一阶模型。

16.1.1 模拟曲线关系

我们能用式（16-1）模拟形式更复杂的关系。为了说明如何做到这一点，我们考虑生产工业天平和实验室设备的 Reynolds 公司所面对的问题。Reynolds 公司的管理人员希望对公司销售人员工作年限的长短和电子实验室天平的销售数量之间的关系进行调研。表 16-1 给出了 15 名随机抽选的销售人员近期天平的销售数量和每一名销售人员被公司雇用的月数。图 16-1 给出了这些数据的散点图。散点图表明，在被公司雇用时间的长短和销售数量之间，可能存在一个曲线关系。在考虑如何为 Reynolds 公司建立一个曲线关系之前，让我们首先考虑在图 16-2 中给出的与简单一阶模型对应的 Minitab 输出；估计的回归方程是

$$\text{Sales} = 111 + 2.38\text{Months}$$

式中，Sales 代表电子实验室天平的销售数量；Months 代表销售人员被公司雇用的月数。

表 16-1 Reynolds 公司例子的数据

销售人员被公司雇用的月数	天平的销售数量
41	275
106	296
76	317
10	376
22	162
12	150
85	367
111	308
40	189
51	235
9	83
12	112
6	67
56	325
19	189

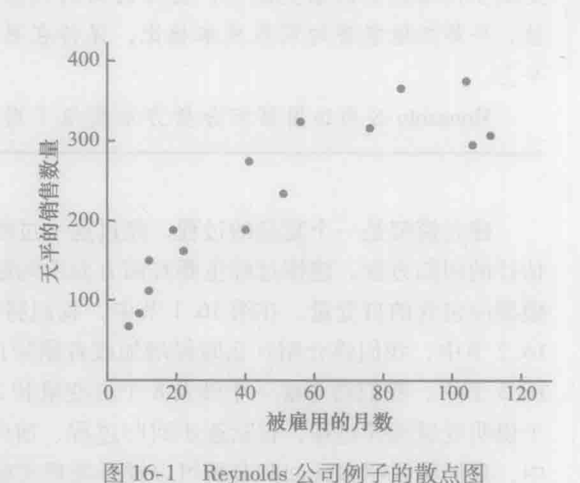


图 16-1 Reynolds 公司例子的散点图

图 16-3 是对应的标准化残差图。尽管计算机输出表明：这个线性关系是显著的（p-值 = 0.000），并且线性关系解释了销售数量中的大部分变异性（$R-Sq = 78.1\%$），然而标准化残差图启发我们，仍然需要一个曲线关系。

为了说明这是一个曲线关系，我们令式（16-1）中的 $z_1 = x_1$，$z_2 = x_1^2$，于是得到模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon \tag{16-3}$$

这个模型被称为具有一个预测变量的二阶模型。为了建立与这个二阶模型相对应的估计的回归方程，我们使用的统计软件包不但需要表 16-1 中的原始数据，而且也需要增加第二个自变量所对应的数据，即销售人员被公司

⊖ 如果你能写出形如式（16-1）的回归模型，那么就能应用在第 15 章中叙述的标准多元回归过程。

雇用月数的平方。在图 16-4 中，我们给出了对应二阶模型的 Minitab 输出；估计的回归方程是

$$\text{Sales} = 45.3 + 6.34\text{Months} - 0.0345\text{MonthsSq}$$

式中，MonthsSq ㊀代表销售人员被公司雇用月数的平方。

```
The regression equation is
Sales = 111 + 2.38 Months

Predictor     Coef   SE Coef     T      p
Constant    111.23     21.63  5.14  0.000
Months      2.3768    0.3489  6.81  0.000

S = 49.5158    R-Sq = 78.1%    R-Sq(adj) = 76.4%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       1  113783  113783  46.41  0.000
Residual Error  13   31874    2452
Total           14  145657
```

图 16-2　Reynolds 公司例子：一阶模型的 Minitab 输出

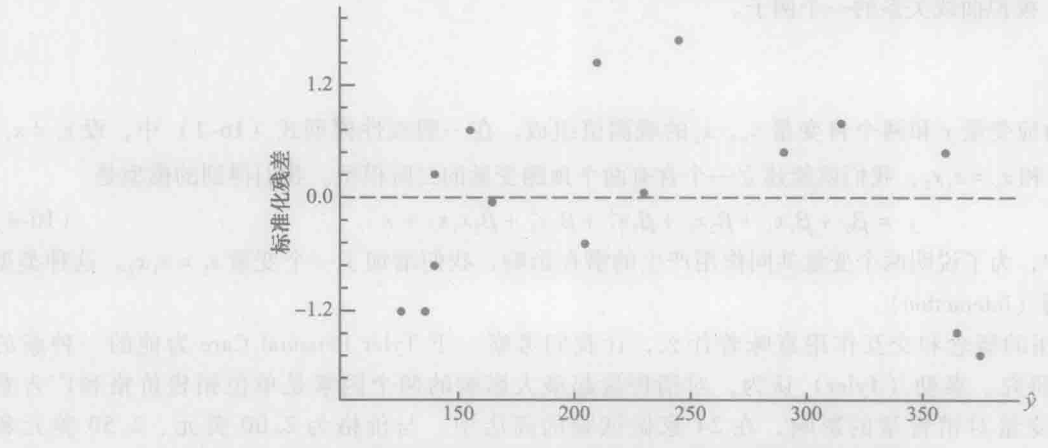


图 16-3　Reynolds 公司例子：一阶模型的标准化残差图

```
The regression equation is
Sales = 45.3 + 6.34 Months - 0.0345 MonthsSq

Predictor        Coef   SE Coef      T      p
Constant        45.35     22.77   1.99  0.070
Months          6.345     1.058   6.00  0.000
MonthsSq    -0.034486  0.008948  -3.85  0.002

S = 34.4528    R-Sq = 90.2%    R-Sq(adj) = 88.6%

Analysis of Variance

SOURCE          DF      SS     MS      F      p
Regression       2  131413  65707  55.36  0.000
Residual Error  12   14244   1187
Total           14  145657
```

图 16-4　Reynolds 公司例子：二阶模型的 Minitab 输出

㊀ 将 Months 的数值平方，就能得到自变量 MonthsSq 的数据。

图 16-5 是对应的标准化残差图。这张图表明：前面图 16-3 中的曲线模式的图形已经被消除。在 $\alpha=0.05$ 的显著性水平下，计算机输出表明，模型在总体上是显著的（F 检验的 p-值是 0.000）；我们还注意到，对应于自变量 MonthsSq 的 t 检验的 p-值（p-值 = 0.002）小于 0.05，因此我们的结论是，在含有自变量 Months 的模型中，增加的自变量 MonthsSq 是显著的。由于 R-Sq（adj）= 88.6%，由这个估计的回归方程给出的拟合我们应该是满意的。不过更重要的是我们看到了，在回归分析中运用曲线关系是多么容易。

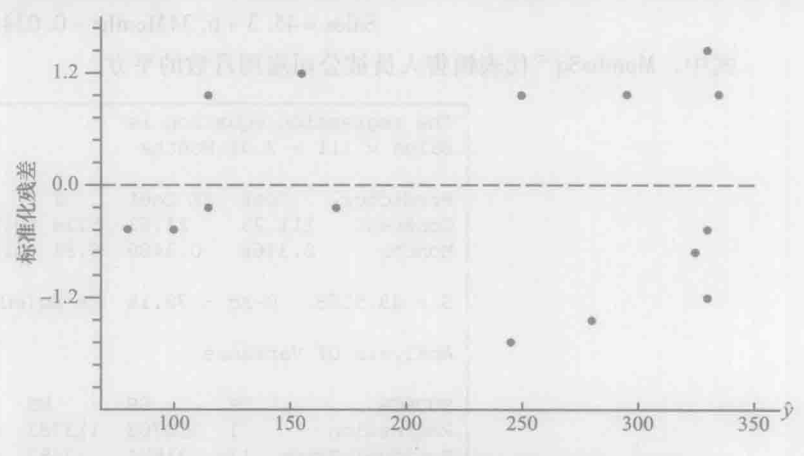


图 16-5 Reynolds 公司例子：二阶模型的标准化残差图

显然，有许多类型的关系能用式（16-1）模拟。我们一直使用的回归方法不一定局限于线性关系或直线关系。在多元回归分析中，线性这个词在术语"一般线性模型"中指的仅仅是这样一个事实：β_0，β_1，…，β_p 全是一次幂，这并不意味着 y 和这些 x_i 之间存在着线性关系。实际上，在这一节中我们已经看到了如何利用式（16-1）模拟曲线关系的一个例子。

16.1.2 交互作用

如果原始数据集由应变量 y 和两个自变量 x_1，x_2 的观测值组成，在一般线性模型式（16-1）中，设 $z_1=x_1$，$z_2=x_2$，$z_3=x_1^2$，$z_4=x_2^2$ 和 $z_5=x_1x_2$，我们就能建立一个含有两个预测变量的二阶模型。我们得到的模型是

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1^2+\beta_4x_2^2+\beta_5x_1x_2+\varepsilon \tag{16-4}$$

在这个二阶模型中，为了说明两个变量共同作用产生的潜在影响，我们增加了一个变量 $z_5=x_1x_2$。这种类型的影响被称为**交互作用**（interaction）。

为了解释交互作用的概念和交互作用意味着什么，让我们考察一下 Tyler Personal Care 为他的一种新的洗发产品进行的回归研究。泰勒（Tyler）认为，对销售量起最大影响的两个因素是单位销售价格和广告费用。为了研究这两个变量对销售量的影响，在 24 家做试验的商店中，与价格为 2.00 美元、2.50 美元和 3.00 美元相对应的广告费用分别为 50 000 美元和 100 000 美元。我们观测到的销售数量（单位：1 000 瓶）记录在表 16-2 中。

表 16-2 Tyler Personal Care 例子的数据

价格（美元）	广告费用（1 000 美元）	销售数量（1 000 瓶）	价格（美元）	广告费用（1 000 美元）	销售数量（1 000 瓶）
2.00	50	478	2.00	100	810
2.50	50	373	2.50	100	653
3.00	50	335	3.00	100	345
2.00	50	473	2.00	100	832
2.50	50	358	2.50	100	641
3.00	50	329	3.00	100	372
2.00	50	456	2.00	100	800
2.50	50	360	2.50	100	620
3.00	50	322	3.00	100	390
2.00	50	437	2.00	100	790
2.50	50	365	2.50	100	670
3.00	50	342	3.00	100	393

表 16-3 是这些数据的汇总。注意，对应销售价格为 2.00 美元、广告费用为 50 000 美元的样本平均销售数量是 461 000 瓶；对应销售价格为 2.00 美元、广告费用为 100 000 美元的样本平均销售数量是 808 000 瓶。因此，当价格为 2.00 美元保持不变时，广告费用分别为 50 000 美元和 100 000 美元，这时候的样本平均销售数量之差是 808 000 − 461 000 = 347 000 瓶。当产品的价格为 2.50 美元时，样本平均销售数量之差是 646 000 − 364 000 = 282 000 瓶。最后，当产品的价格为 3.00 美元时，样本平均销售数量之差是 375 000 − 332 000 = 43 000 瓶。显然，广告费用分别为 50 000 美元和 100 000 美元时，样本平均销售数量之差依赖于产品的销售价格。换句话说，当销售价格较高时，增加广告费用的影响将要减少。上述观测结果提供了销售价格和广告费用这两个变量之间交互作用的证据。

表 16-3　Tyler Personal Care 例子的样本平均销售数量（1 000 瓶）

		价格（美元）		
		2.00	2.50	3.00
广告费用（美元）	50 000	461	364	332
	100 000	808①	646	375

①当价格为 2.00 美元和广告费用为 100 000 美元时，平均销售数量为 808 000 瓶。

为了给出交互作用的另一个观点，图 16-6 表示了 6 种不同的销售价格与广告费用组合的样本平均销售数量。该图还表明了广告费用对样本平均销售数量的影响依赖于产品的价格水平，这让我们再一次看到了交互作用的影响。当两个变量之间存在交互作用时，我们不可能独立于其他变量之外，单独地考察一个变量对响应变量 y 的影响。换句话说，只有当我们考虑两个变量对响应变量的联合影响时，才能得出有意义的结论。

为了说明交互作用的影响，我们将利用下面的回归模型。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \quad (16\text{-}5)$$

式中，y 为销售数量（1 000 瓶）；x_1 为销售价格（美元）；x_2 为广告费用（1 000 美元）。

注意，式（16-5）反映出 Tyler 相信销售数量线性地依赖销售价格和广告费用（由 $\beta_1 x_1$ 和 $\beta_2 x_2$ 两项说明），并且两个变量之间存在交互作用（由 $\beta_3 x_1 x_2$ 项说明）。

为了建立估计的回归方程，我们利用含有 3 个自变量（z_1，z_2 和 z_3）的一般线性模型。

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \varepsilon \quad (16\text{-}6)$$

式中，z_1 取代 x_1；z_2 取代 x_2；z_3 取代 $x_1 x_2$。

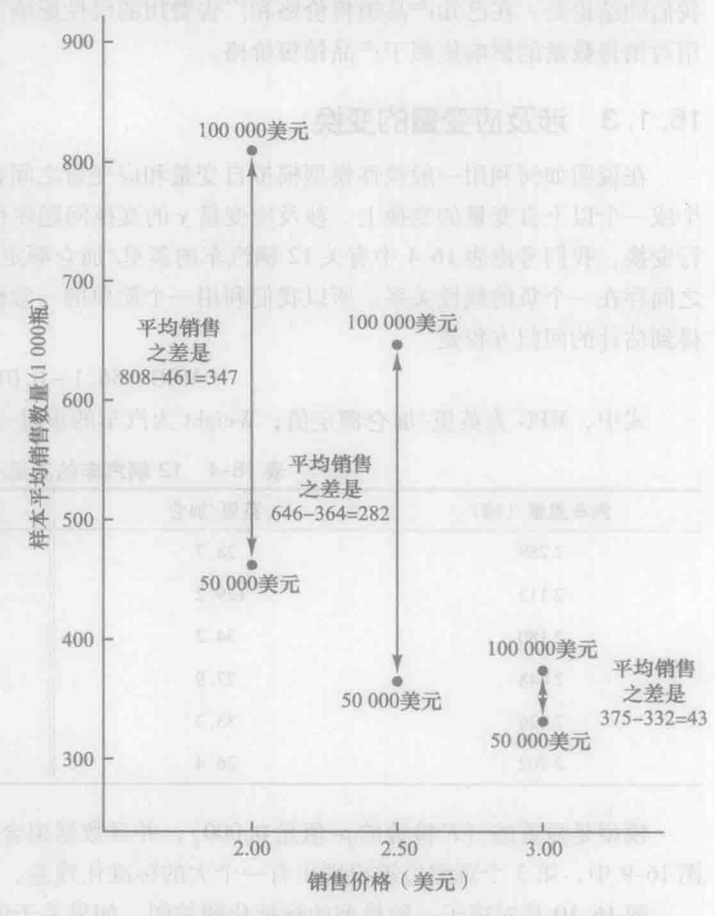


图 16-6　样本平均销售数量是销售价格和广告费用的函数

图 16-7 是对应于 Tyler Personal Care 例子的交互作用模型的 Minitab 输出。得到估计的回归方程是

Sales = −276 + 175Price + 19.7AdvExp − 6.08PriceAdv

式中，Sales 代表销售数量（1 000 瓶）；Price 代表销售价格（美元）；AdvExp 代表广告费用（1 000 美元）；PriceAdv ㊀ 代表交互作用项（Price × AdvExp）。

㊀ 将 Price 的每一个数值与对应的 AdvExp 的数值相乘，就能得到自变量 PriceAdv 的数据。

```
The regression equation is
Sales = - 276 + 175 Price + 19.7 AdvExpen - 6.08 PriceAdv

Predictor      Coef   SE Coef       T      p
Constant     -275.8    112.8    -2.44  0.024
Price        175.00    44.55     3.93  0.001
Adver        19.680    1.427    13.79  0.000
PriceAdv    -6.0800   0.5635   -10.79  0.000

S = 28.1739   R-Sq = 97.8%   R-Sq(adj) = 97.5%

Analysis of Variance

SOURCE          DF      SS      MS       F      p
Regression       3  709316  236439  297.87  0.000
Residual Error  20   15875     794
Total           23  725191
```

图 16-7 Tyler Personal Care 例子的 Minitab 输出

因为模型在总体上是显著的（F 检验的 p-值是 0.000），对应于交互作用项 PriceAdv 的 t 检验的 p-值是 0.000，我们的结论是：在已知产品销售价格和广告费用的线性影响下，交互作用是显著的。于是回归结果表明，广告费用对销售数量的影响依赖于产品销售价格。

16.1.3 涉及应变量的变换

在说明如何利用一般线性模型模拟自变量和应变量之间各种可能存在的关系时，我们把注意力集中到包含一个或一个以上自变量的变换上。涉及应变量 y 的变换问题往往是值得考虑的。为了解释什么时候应该对应变量进行变换，我们考虑表 16-4 中有关 12 辆汽车的英里/加仑额定值和重量的数据。散点图 16-8 表明了在这两个变量之间存在一个负的线性关系。所以我们利用一个简单的一阶模型把这两个变量联系起来。图 16-9 是 Minitab 输出；得到估计的回归方程是

$$\text{MPG} = 56.1 - 0.0116\text{Weight}$$

式中，MPG 为英里/加仑额定值；Weight 为汽车的重量（磅）。

表 16-4 12 辆汽车的英里/加仑额定值和重量

汽车重量（磅）	英里/加仑	汽车重量（磅）	英里/加仑
2 289	28.7	2 657	23.9
2 113	129.2	2 106	30.5
2 180	34.2	3 226	18.1
2 448	27.9	3 213	19.5
2 026	33.3	3 607	14.3
2 702	26.4	2 888	20.9

模型是显著的（F 检验的 p-值是 0.000），并且数据拟合得也非常好（R-Sq = 93.5%）。然而，我们注意到在图 16-9 中，第 3 个观测值被识别出有一个大的标准化残差。

图 16-10 是对应于一阶模型的标准化残差图。如果关于误差项的假设全都成立，我们期望看到的应是一条水平的带状图形，然而我们观察到的图形看起来不像是这种形状。实际上，残差的变异性看来随着$\hat{y}$值的增加而增加。换句话说，我们看到的是楔形图，而这种图形在第 14 章和第 15 章被认为是存在非常数方差的征兆。当显著性检验的基本假设看起来没有被满足时，我们不能证明：得到的有关估计的回归方程统计显著性的任何结论都是合理的。

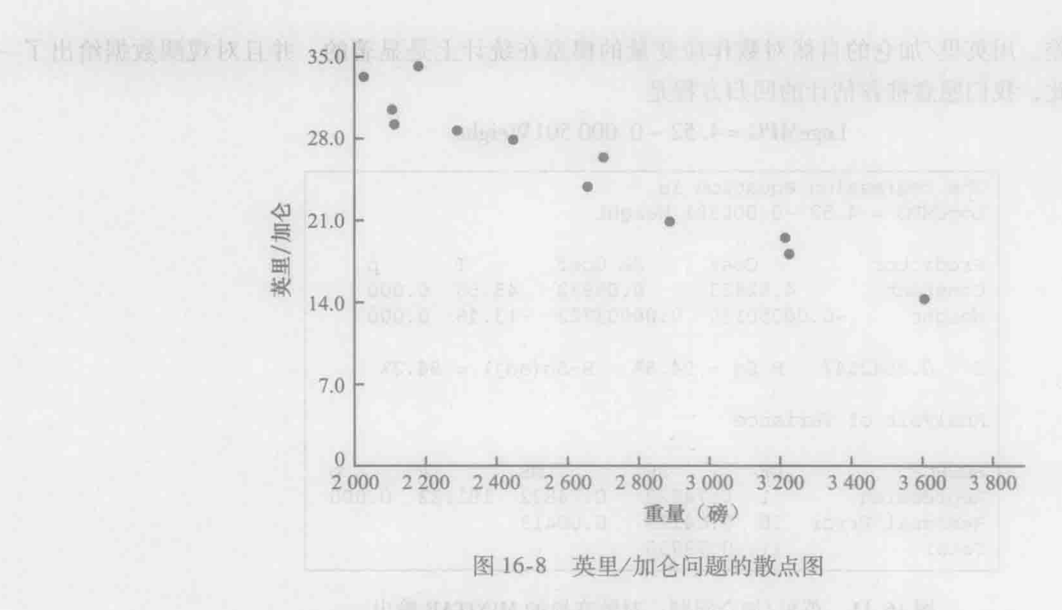
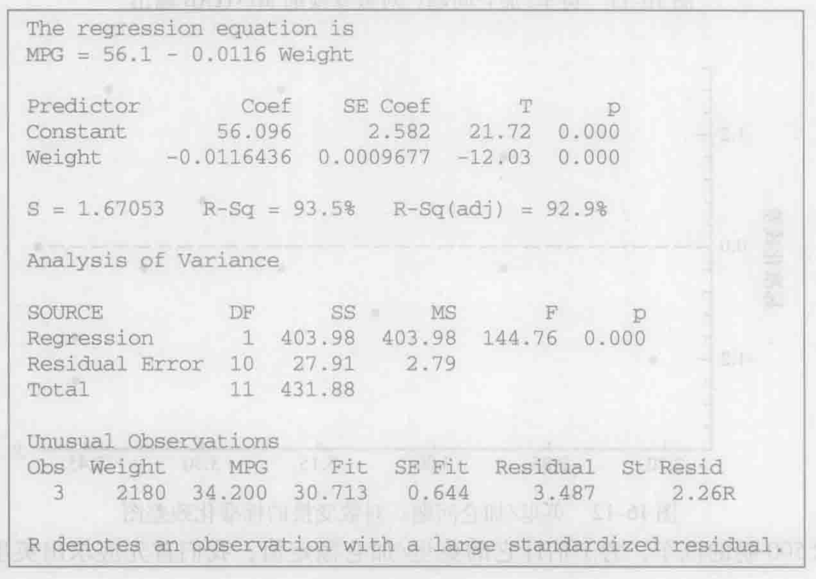


图 16-8 英里/加仑问题的散点图



```
The regression equation is
MPG = 56.1 - 0.0116 Weight

Predictor        Coef     SE Coef       T      p
Constant       56.096       2.582   21.72  0.000
Weight     -0.0116436   0.0009677  -12.03  0.000

S = 1.67053   R-Sq = 93.5%   R-Sq(adj) = 92.9%

Analysis of Variance

SOURCE          DF      SS      MS       F      p
Regression       1  403.98  403.98  144.76  0.000
Residual Error  10   27.91    2.79
Total           11  431.88

Unusual Observations
Obs  Weight     MPG     Fit  SE Fit  Residual  St Resid
  3    2180  34.200  30.713   0.644     3.487     2.26R

R denotes an observation with a large standardized residual.
```

图 16-9 英里/加仑问题的 Minitab 输出

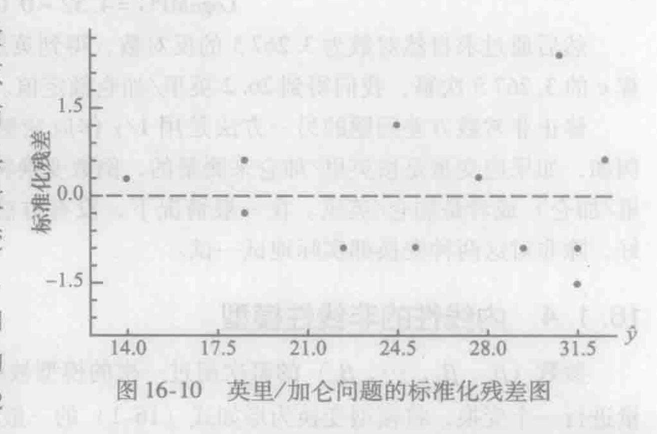


图 16-10 英里/加仑问题的标准化残差图

非常数方差问题通常能够被修正，修正的方法是对应变量做一个不同的比例变换。例如，用应变量的对数来代替原来的应变量，这样做的效果是压缩了应变量的数值，从而达到减少非常数方差影响的目的。大部分统计软件包都具有对数变换的功能，或者以 10 为底（常用对数）进行对数变换，或者以 e = 2.718 28…为底（自然对数）进行对数变换。我们对英里/加仑问题的原始数据进行自然对数变换，并且建立英里/加仑的自然对数关于汽车重量的估计的回归方程。我们用英里/加仑的自然对数作应变量，在输出中的标记为 LogeMPG，得到的回归结果在图 16-11 中；图 16-12 是对应的标准化残差图。

我们注意一下图 16-12 中的标准化残差图，现在我们已经看不到楔形图了。此外，没有一个观测值被识别出

有一个大的标准化残差。用英里/加仑的自然对数作应变量的模型在统计上是显著的，并且对观测数据给出了一个非常好的拟合。因此，我们愿意推荐估计的回归方程是

$$\text{LogeMPG} = 4.52 - 0.000\,501\,\text{Weight}$$

```
The regression equation is
LogeMPG = 4.52 -0.000501 Weight

Predictor          Coef      SE Coef        T       p
Constant        4.52423      0.09932    45.55   0.000
Weight      -0.00050110   0.00003722   -13.46   0.000

S = 0.0642547   R-Sq = 94.8%   R-Sq(adj) = 94.2%

Analysis of Variance

SOURCE          DF       SS        MS        F       p
Regression       1  0.74822   0.74822   181.22   0.000
Residual Error  10  0.04129   0.00413
Total           11  0.78950
```

图 16-11 英里/加仑问题：对数变换的 MINITAB 输出

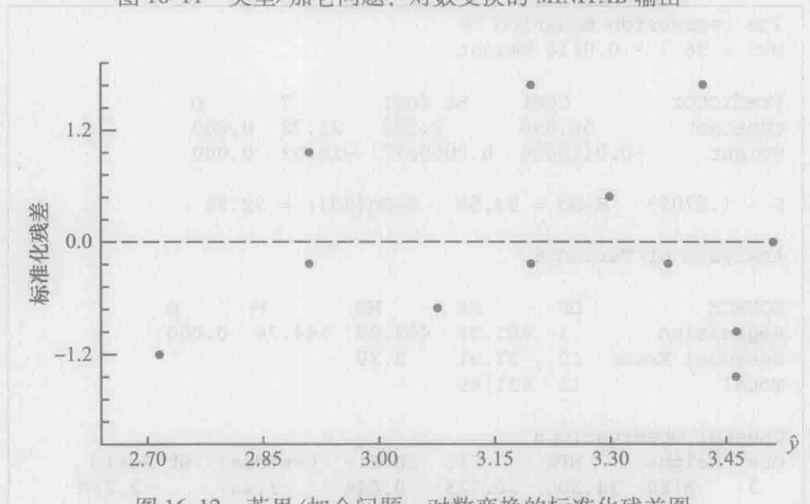


图 16-12 英里/加仑问题：对数变换的标准化残差图

对于一辆重量为 2 500 磅的汽车，为了估计它的英里/加仑额定值，我们首先应求出英里/加仑额定值的自然对数的估计值。

$$\text{LogeMPG} = 4.52 - 0.000\,501 \times 2\,500 = 3.267\,5$$

然后通过求自然对数为 3.267 5 的反对数，得到英里/加仑额定值的估计值。利用具有指数功能的计算器，计算 e 的 3.267 5 次幂，我们得到 26.2 英里/加仑额定值。

修正非常数方差问题的另一方法是用 $1/y$ 作应变量来代替原来的应变量 y。这种类型的变换叫做倒数变换。例如，如果应变量是按英里/加仑来测量的，倒数变换将得到一个新的应变量，这个新的应变量的单位是 1/（英里/加仑）或者是加仑/英里。在一般情况下，没有方法能决定，究竟是进行对数变换还是进行倒数变换效果更好，除非对这两种变换都实际地试一试。

16.1.4 内线性的非线性模型

参数（β_0，β_1，…，β_p）的幂次超过一次的模型被称为非线性模型。然而，对于指数模型情形，我们能对变量进行一个变换，将模型变换为形如式（16-1）的一般线性模型，从而使得我们能进行回归分析。指数模型与下面的回归方程有关。

$$E(y) = \beta_0 \beta_1^x \tag{16-7}$$

当应变量 y 随着 x 的增加，按一个不变的百分比，而不是一个固定的数量增加或减少时，适合应用这种回归模型。

作为一个例子，假设一种产品的销售收入 y 依赖于广告费用 x(1 000 美元)，这个问题对应的指数模型如下所示。

$$E(y) = 500 \times (1.2)^x$$

于是，对于 $x=1$，$E(y)=500\times(1.2)^1=600$；对于 $x=2$，$E(y)=500\times(1.2)^2=720$；对于 $x=3$，$E(y)=500\times(1.2)^3=864$。注意，在这种情形下，$E(y)$ 不是按一个固定的数量增加，而是按一个不变的百分比增加；增长的百分比是20%。

我们能通过对式（16-7）两边取对数，将这个非线性回归方程转化为一个线性回归方程。

$$\log E(y) = \log\beta_0 + x\log\beta_1 \qquad (16\text{-}8)$$

注意，如果我们设 $y'=\log E(y)$，$\beta_0'=\log\beta_0$，$\beta_1'=\log\beta_1$，我们能将式（16-8）写成

$$y' = \beta_0' + \beta_1' x$$

显然，我们能利用简单线性回归的公式求出 β_0'和 β_1'的估计量。用 b_0'和 b_1'表示估计量，得到下面估计的回归方程。

$$\hat{y}' = b_0' + b_1' x \qquad (16\text{-}9)$$

为了在已知 x 的值时，得到原来的应变量 y 的预测值，我们首先将 x 的值代入式（16-9）中并计算出$\hat{y}'$。$\hat{y}'$的反对数就是我们要求的 y 的预测值，或者是 y 的期望值。

许多非线性模型不能被转化为一个等价的线性模型。所以这种模型限制了我们在商务和经济中的应用。此外，研究这种模型所必需的数学背景超出了这本教科书的范围。

练　习

方法

2. 考虑两个变量 x 和 y 的以下数据。

x	9	32	18	15	26
y	10	20	21	16	22

a. 对这些数据建立一个形如$\hat{y}=b_0+b_1x$ 的估计的回归方程。对这个方程是否适宜对 y 的值进行预测做出评述。

b. 对这些数据建立一个形如$\hat{y}=b_0+b_1x+b_2x^2$的估计的回归方程。对这个方程是否适宜对 y 的值进行预测做出评述。

c. 当 $x=20$ 时，预测 y 的值。

应用

4. 公路管理部门进行一项有关交通流量和车速之间关系的研究。假设模型的形式如下所示。

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

式中，y 代表交通流量（辆/小时）；x 代表车速（英里/小时）。

下面是在下班高峰期间对 6 个主要出城交通道口采集的数据。

交通流量（y）	车速（x）	交通流量（y）	车速（x）
1 256	35	1 335	45
1 329	40	1 349	50
1 226	30	1 124	25

a. 对这些数据建立一个估计的回归方程。

b. 在 $\alpha=0.01$ 的显著性水平下，检验 x 和 y 之间的显著关系。

6. 在一项有关应急服务交通工具的研究中，研究人员要调查提供应急服务的交通工具的数量和行驶的平均距离之间的关系。下表给出了采集的数据。

应急服务交通工具的数量	行驶的平均距离（英里）
9	1.66
11	1.12
16	0.83
21	0.62
27	0.51
30	0.47

a. 将处理应急服务交通工具需要行驶的平均距离视

为应变量，对这些数据画出散点图。

b. 简单线性回归模型是否是一个适宜的模型？请做出解释。

c. 对这些数据，建立一个你认为能很好地解释这两个变量之间关系的估计的回归方程。

8. 克尔维特、法拉利和捷豹生产了各式各样的持续升值的老爷车。根据马丁汽车收藏评估系统（Martin Rating System for Collectible Cars），下面是15辆老爷车的稀有等级（1~20）和价格（单位：1 000美元）的统计数据（*BusinessWeek* website，2006年2月）。

a. 以稀有等级为自变量，以价格为应变量，画出这些数据的散点图。简单线性回归模型是一个合适的模型吗？

b. 令 x = 稀有等级，以 x 和 x^2 做为两个自变量，建立一个估计的多元回归方程。

c. 参考由式（16-7）给出的非线性关系。对这个模型，利用对数变换建立一个估计的回归方程。

d. 你推荐的估计的回归方程是在（b）中建立的还是在（c）中建立的？请做出解释。

年份	品牌	型号	等级	价格（1 000美元）
1984	雪佛兰	克尔维特	18	1 600.0
1956	雪佛兰	克尔维特265/225-hp	19	4 000.0
1963	雪佛兰	克尔维特coupe（340-bhp 4-speed）	18	1 000.0
1978	雪佛兰	克尔维特coupe Silver Anniversary	19	1 300.0
1960~1963	法拉利	250 GTE 2+2	16	350.0
1962~1964	法拉利	250 GTLLusso	19	2 650.0
1962	法拉利	250 GTO	18	375.0
1967~1968	法拉利	275 GTB/4 NARTSpyder	17	450.0
1968~1973	法拉利	365 GTB/4 Daytona	17	140.0
1962~1967	捷豹	E-type OTS	15	77.5
1969~1971	捷豹	E-type Series Ⅱ OTS	14	62.0
1971~1974	捷豹	E-type Series Ⅲ OTS	16	125.0
1951~1954	捷豹	XK 120 roadster（steel）	17	400.0
1950~1953	捷豹	XK C-type	16	250.0
1956~1957	捷豹	XKSS	13	70.0

16.2 确定什么时候增加或者删除变量

在这一节我们将说明，如何利用 F 检验来确定，将一个或一个以上的自变量增加到一个多元回归模型上是否适宜的问题。这一检验的根据是：测定一个多元回归模型增加一个或一个以上的自变量所得到的误差平方和减少的数量。首先我们将说明，怎样将这一检验应用到上一章介绍的Butler运输公司的例子中去。

在第15章，我们用Butler运输公司的例子来说明多元回归分析的应用。我们回想一下，公司的管理人员希望利用运输车辆每天运送货物的行驶里程和运送货物的次数这两个自变量建立一个估计的回归模型，并用这个模型去预测运输车辆全天的行驶时间。只用运输车辆运送货物的行驶里程 x_1 作自变量，利用最小二乘法得到下面的估计的回归方程。

$$\hat{y} = 1.27 + 0.0678x_1$$

在第15章我们已经给出了这个模型的误差平方和是SSE = 8.029。当运送货物的次数 x_2 作为第二个自变量增加到模型上时，我们得到下面的估计的回归方程。

$$\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$$

这个模型的误差平方和是SSE = 2.299。显然，增加 x_2 导致SSE的数量减少。我们希望回答的问题是：增加变量 x_2 是否导致了SSE的数量显著地减少？

当模型中 x_1 是唯一的自变量时，我们用记号SSE(x_1) 表示模型的误差平方和；当模型中有两个自变量 x_1，x_2

时，我们用记号SSE(x_1，x_2）表示模型的误差平方和，等等。因此，在仅包含自变量x_1的模型中，将x_2增加到模型上，引起SSE减少的数量是

$$\text{SSE}(x_1) - \text{SSE}(x_1, x_2) = 8.029 - 2.299 = 5.730$$

我们进行F检验去确定，这一数量上的减少是否是显著的。

F统计量的分子是用SSE减少的数量除以增加到原模型上的自变量的个数。这里，仅有一个自变量x_2增加到模型上，于是F统计量的分子是

$$\frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{1} = 5.730$$

得到的这个结果是模型每增加一个自变量，误差平方和SSE减少数量的度量。F统计量的分母是包括全部自变量的模型的均方误差。对于Butler运输公司的例子，对应的模型含有两个自变量x_1和x_2，于是$p=2$，并且有

$$\text{MSE} = \frac{\text{SSE}(x_1, x_2)}{n-p-1} = \frac{2.299}{7} = 0.3284$$

下面的F统计量给出了将自变量x_2增加到模型上，在统计上是否显著的检验。

$$F = \frac{\dfrac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{1}}{\dfrac{\text{SSE}(x_1, x_2)}{n-p-1}} \tag{16-10}$$

这个F检验的分子自由度等于增加到模型中的自变量的个数，分母自由度等于$n-p-1$。

对于Butler运输公司的例子，我们得到

$$F = \frac{\dfrac{5.730}{1}}{\dfrac{2.299}{7}} = \frac{5.730}{0.3284} = 17.45$$

查阅附录B的表B-4，对于显著性水平$\alpha=0.05$，我们得到$F_{0.05}=5.59$。因为$F=17.45>F_{0.05}=5.59$，于是我们应该拒绝x_2在统计上是不显著的原假设；换句话说，在仅含有自变量x_1的模型中，增加自变量x_2到模型上，将引起误差平方和显著地减少。

当我们想要检验仅增加一个自变量到模型上的显著性时，刚刚叙述的F检验得到的结果，也能通过使用在第15.4节中介绍的对单个参数的t检验得到。实际上，刚刚才计算的F统计量是用来检验单个参数显著性的t统计量的平方。

因为当模型仅增加一个自变量时，t检验与F检验是等价的，所以我们现在能进一步说明，为了检验单个参数的显著性，可以适当地利用t检验。如果单个参数是不显著的，那么对应的变量就能从模型中删除。但是，如果t检验显示，有两个或两个以上的参数是不显著的，那么基于t检验，从模型中删除的自变量永远不能超过一个；如果有一个变量从模型中被删除，那么最初不显著的第二个变量可能变成一个显著的变量。

我们现在着手考虑增加一个以上的自变量，即一组变量时，是否会引起误差平方和有一个显著减少的问题。

16.2.1　一般情形

考虑以下含有q个自变量的多元回归模型，这里$q<p$。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon \tag{16-11}$$

如果增加自变量x_{q+1}，x_{q+2}，…，x_p到这个模型上，我们就得到一个含有p个自变量的多元回归模型。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \beta_{q+2} x_{q+2} + \cdots + \beta_p x_p + \varepsilon \tag{16-12}$$

为了检验增加的自变量x_{q+1}，x_{q+2}，…，x_p是否在统计上是显著的，我们提出的原假设和备择假设叙述如下。

$$H_0: \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$$

$$H_a: 参数\ \beta_{q+1}, \beta_{q+2}, \cdots, \beta_p\ 中至少有一个不等于零$$

下面的 F 统计量给出了检验增加的自变量 x_{q+1}，x_{q+2}，…，x_p在统计上是否显著的根据。

$$F = \frac{\dfrac{SSE(x_1, x_2, \cdots, x_q) - SSE(x_1, x_2, \cdots, x_q, x_{q+1} \cdots, x_p)}{p - q}}{\dfrac{SSE(x_1, x_2, \cdots, x_q, x_{q+1}, \cdots, x_p)}{n - p - 1}} \tag{16-13}$$

然后，将计算出的 F 统计量的值与分子的自由度为 $p-q$、分母的自由度为 $n-p-1$ 的 F 分布表的上侧分位数 F_α 进行比较。如果 $F > F_\alpha$，我们拒绝 H_0，并且可以得出结论：增加的这组自变量在统计上是显著的。注意，对于 $q=1$，$p=2$ 这种特殊情形，式（16-13）就简化为式（16-10）。[⊖]

许多学生感到式（16-13）有些复杂。为了给出 F 统计量一个比较简单的表述，我们将自变量个数较少的模型称为简化模型，将自变量个数较多的模型称为完全模型。如果我们用 SSE（简化）表示简化模型的误差平方和，用 SSE（完全）表示完全模型的误差平方和，我们能把式（16-13）的分子写成

$$\frac{SSE(简化) - SSE(完全)}{增加的项数} \tag{16-14}$$

注意，式（16-14）中的分母“增加的项数”表示完全模型的自变量个数和简化模型的自变量个数之间的差。式（16-13）的分母是完全模型的误差平方和除以对应的自由度；换句话说，式（16-13）的分母是完全模型的均方误差。我们用 MSE（完全）表示完全模型的均方误差，于是我们就能将式（16-13）写成

$$F = \frac{\dfrac{SSE(简化) - SSE(完全)}{增加的项数}}{MSE(完全)} \tag{16-15}$$

为了说明这个 F 统计量的应用，假设我们有一个含有 30 个观测值的回归问题。第一个模型的自变量是 x_1，x_2 和 x_3，它的误差平方和为 150；第二个模型的自变量是 x_1，x_2，x_3，x_4 和 x_5，它的误差平方和为 100。增加两个自变量 x_4 和 x_5 到第一个模型上，会引起误差平方和显著地减少吗？

首先，我们注意到：SST 的自由度是 $30-1=29$，完全模型的回归平方和的自由度是 5（在完全模型中自变量的个数）。于是，完全模型的误差平方和的自由度是 $29-5=24$，因此 MSE（完全）$=100/24=4.17$。所以 F 统计量是

$$F = \frac{\dfrac{150 - 100}{2}}{4.17} = 6.00$$

将这个计算出来的 F 统计量的值与分子的自由度为 2，分母的自由度为 24 的 F 分布表的值进行比较。在 $\alpha = 0.05$ 的显著性水平下，由附录 B 表 B-4，我们查出 $F_{0.05}=3.40$。因为 $F(=6.00) > F_{0.05}(=3.40)$，我们的结论是：增加的自变量 x_4 和 x_5 在统计上是显著的。

16.2.2 p-值的应用

我们还能利用 p-值准则来确定，增加一个或一个以上的自变量到一个多元回归模型上是否适宜的问题。在上面的例子中我们已经说明，如果将两个自变量 x_4 和 x_5 增加到有 3 个自变量 x_1，x_2 和 x_3 的模型上时，怎样实施 F 检验来确定，这一增加在统计上是显著的。对于这个例子，计算出的 F 统计量的值是 6.00，将 $F=6.00$ 与临界值 $F_{0.05}=3.40$ 进行比较，我们的结论是：增加的自变量 x_4 和 x_5 在统计上是显著的。与分子的自由度为 2，分母的自由度为 24 的 F 统计量的值 $F=6.00$ 相联系的 p-值 $=0.008<0.05$，于是我们也能得出结论，增加的两个自变量 x_4 和 x_5 在统计上是显著的。从 F 分布表中直接查出 p-值是一件困难的事，但是计算机软件包，例如 Minitab 或 Excel 可以不费力地计算出 p-值。

[⊖] 许多计算机软件包，例如 Minitab，给出了按顺序进入模型的每个自变量所对应的增加的平方和；在这种情形下，确定是否增加或者删除一组变量的 F 检验的计算被简化了。

注释和评论

F统计量的计算也能根据回归平方和的差来完成。为了说明这种计算形式的F统计量，首先，我们注意到

$$SSE(\text{简化}) = SST - SSR(\text{简化})$$
$$SSE(\text{完全}) = SST - SSR(\text{完全})$$

所以

$$SSE(\text{简化}) - SSE(\text{完全}) = [SST - SSR(\text{简化})] - [SST - SSR(\text{完全})] = SSR(\text{完全}) - SSR(\text{简化})$$

于是

$$F = \frac{\dfrac{SSR(\text{完全}) - SSR(\text{简化})}{\text{自变量增加的个数}}}{MSE(\text{完全})}$$

练习

方法

10. 在含有27个观测值的回归分析中，我们已经建立了下面的估计的回归方程。

$$\hat{y} = 25.2 + 5.5x_1$$

对于这个估计的回归方程，SST = 1 550，SSE = 520。

a. 在$\alpha = 0.05$的显著性水平下，检验x_1是否是显著的。

假设将自变量x_2和x_3增加到模型上，我们得到了下面的估计的回归方程。

$$\hat{y} = 16.3 + 2.3x_1 + 12.1x_2 - 5.8x_3$$

对于这个估计的回归方程，SST = 1 550，SSE = 100。

b. 在$\alpha = 0.05$的显著性水平下，利用F检验去确定，x_2和x_3对模型的显著性是否有作用。

应用

12. 美国女子职业高尔夫球协会保存了高尔夫球员参加LPGA巡回赛的成绩和奖金的统计资料。在2005年LPGA巡回赛的赛事中，总奖金前30位的高尔夫球员的年终成绩的统计资料，保存在本书所附光盘名为LPGATour的文件中（LPGA Tour website，2006年）。有关数据的描述如下所示。

Earnings（$1 000）是以1 000美元为单位的总奖金；

ScoringAvg. 是参加全部赛事的平均击球次数；

Greens in Reg. 是高尔夫球员能够标准杆上果岭次数的比例；

PuttingAvg. 是将标准杆上果岭的高尔夫球推杆入洞的平均杆数；

Sand Saves 是一旦高尔夫球员将高尔夫球打入果岭边的沙坑内，高尔夫球员能克服"地面的高低起伏"将球救出的比例。

标准杆上果岭是指，如果高尔夫球的任一部分触及果岭的推杆区域，并且使高尔夫球到达推杆区的杆数与标准杆的差至少是2杆。

a. 建立一个估计的回归方程，使这个方程在高尔夫球员将标准杆上果岭的高尔夫球推杆入洞的平均杆数已知时，能用来预测高尔夫球员参加全部赛事的平均击球次数。

b. 建立一个估计的回归方程，使这个方程在高尔夫球员能够标准杆上果岭次数的比例，将标准杆上果岭的高尔夫球推杆入洞的平均杆数，以及一旦将高尔夫球打入果岭边的沙坑内，高尔夫球员能克服"地面的高低起伏"将球救出的比例已知时，能用来预测高尔夫球员参加全部赛事的平均击球次数。

c. 在$\alpha = 0.05$的显著性水平下，检验在（b）中增加的两个自变量：能够标准杆上果岭次数的比例和一旦将高尔夫球打入果岭边的沙坑内，高尔夫球员能克服"地面的高低起伏"将球救出的比例，对在（a）中建立的估计的回归方程，是否有显著的作用。请作出解释。

14. 美国心脏协会经过十年的研究，得到了有关中风风险与年龄、血压和吸烟嗜好相关性的统计资料。这项研究的部分数据如下表所示。中风风险被认为是一个人在未来的十年内发生中风的概率（乘上100）。对于吸烟嗜好变量，1表示是一名吸烟

者，而0表示是一名不吸烟者。

中风风险	年龄	血压	吸烟嗜好
12	57	152	0
24	67	163	0
13	58	155	0
56	86	177	1
28	59	196	0
51	76	189	1
18	56	155	1
31	78	120	0
37	80	135	1
15	78	98	0
22	71	152	0
36	70	173	1
15	67	135	1
48	77	209	1
15	60	199	0
36	82	119	1

（续）

中风风险	年龄	血压	吸烟嗜好
8	66	166	0
34	80	125	1
3	62	117	0
37	59	207	1

a. 建立一个估计的回归方程，使这个方程在年龄和血压已知时，能用来预测中风的风险。

b. 考虑增加两个自变量到（a）中所建立的模型上，一个自变量是年龄和血压之间的交互作用，另一个自变量是一个人是否有吸烟嗜好。利用这四个自变量建立一个估计的回归方程。

c. 在$\alpha=0.05$的显著性水平下，通过检验是否可以看出，增加交互作用和吸烟嗜好这两个自变量，对在（a）中建立的估计的回归方程有显著的作用。

16.3 大型问题的分析

在介绍多元回归分析时，我们广泛地应用了Butler运输公司的例子。在探讨介绍性的概念时，这个例子涉及的变量较少是一大优点，但是为了说明在建模过程中我们应该选择哪些变量进入模型，这个例子就显得困难了。为了给出将要在下一节详细论述的变量选择过程，我们引入由8个自变量、25组观测值组成的一个数据集。得克萨斯州基督教大学市场营销系的David W. Cravens博士提供了这些数据，并允许我们使用这些数据。所以，我们把这个数据集称为Cravens数据。[⊖]

Cravens数据是一家公司的数据，这家公司在一些销售区域销售产品，并且为每一个销售区域分别指定了独家经销商。进行回归分析的目的是为了确定：预测（自）变量的变化是否能解释每一个销售区域的销售情况。由25个销售区域组成了一个随机样本，得到的数据如表16-5所示；变量的定义在表16-6中给出。

表16-5 Cravens数据

Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work	Rating
3 669.88	43.10	74 065.1	4 582.9	2.51	0.34	74.86	15.05	4.9
3 473.95	108.13	58 117.3	5 539.8	5.51	0.15	107.32	19.97	5.1
2 295.10	13.82	21 118.5	2 950.4	10.91	−0.72	96.75	17.34	2.9
4 675.56	186.18	68 521.3	2 243.1	8.27	0.17	195.12	13.40	3.4
6 125.96	161.79	57 805.1	7 747.1	9.15	0.50	180.44	17.64	4.6
2 134.94	8.94	37 806.9	402.4	5.51	0.15	104.88	16.22	4.5
5 031.66	365.04	50 935.3	3 140.6	8.54	0.55	256.10	18.80	4.6
3 367.45	220.32	35 602.1	2 086.2	7.07	−0.49	126.83	19.86	2.3
6 519.45	127.64	46 176.8	8 846.2	12.54	1.24	203.25	17.42	4.9
4 876.37	105.69	42 053.2	5 673.1	8.85	0.31	119.51	21.41	2.8
2 468.27	57.72	36 829.7	2 761.8	5.38	0.37	116.26	16.32	3.1
2 533.31	23.58	33 612.7	1 991.8	5.43	−0.65	142.28	14.51	4.2
2 408.11	13.82	21 412.8	1 971.5	8.48	0.64	89.43	19.35	4.3
2 337.38	13.82	20 416.9	1 737.4	7.80	1.01	84.55	20.02	4.2
4 586.95	86.99	36 272.0	10 694.2	10.34	0.11	119.51	15.26	5.5

⊖ 详细内容参见David W. Cravens，Robert B. Woodruff 和 Joe C. Stamper，“An Analytical Approach for Evaluating Sales Territory Performance”，*Journal of Marketing*，36（January 1972）：31-37。Copyright 1972 American Marketing Association。

（续）

Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work	Rating
2 729. 24	165. 85	23 093. 3	8 618. 6	5. 15	0. 04	80. 49	15. 87	3. 6
3 289. 40	116. 26	26 878. 6	7 747. 9	6. 64	0. 68	136. 58	7. 81	3. 4
2 800. 78	42. 28	39 572. 0	4 565. 8	5. 45	0. 66	78. 86	16. 00	4. 2
3 264. 20	52. 84	51 866. 1	6 022. 7	6. 31	−0. 10	136. 58	17. 44	3. 6
3 453. 62	165. 04	58 749. 8	3 721. 1	6. 35	−0. 03	138. 21	17. 98	3. 1
1 741. 45	10. 57	23 990. 8	861. 0	7. 37	−1. 63	75. 61	20. 99	1. 6
2 035. 75	13. 82	25 694. 9	3 571. 5	8. 39	−0. 43	102. 44	21. 66	3. 4
1 578. 00	8. 13	23 736. 3	2 845. 5	5. 15	0. 04	76. 42	21. 46	2. 7
4 167. 44	58. 44	34 314. 3	5 060. 1	12. 88	0. 22	136. 58	24. 78	2. 8
2 799. 97	21. 14	22 809. 5	3 552. 0	9. 14	−0. 74	88. 62	24. 96	3. 9

表 16-6　Cravens 数据的变量定义

变量名称	定　义
Sales	经销商总的信贷销售收入
Time	按月计算的经销商为公司从事销售业务的时间
Poten	市场潜力；销售区域总销售数量*
AdvExp	销售区域广告费用
Share	市场份额；过去四年的加权平均
Change	过去四年间市场份额的变化
Accounts	经销商指定的顾客可以赊购的商店数目*
Work	工作量；根据商店一年的零售和批发额计算的加权指数
Rating	根据八个方面的业绩对经销商综合排序；综合排序得分为 1 ~ 7 分

注：*为保护商业秘密，这些数据已被编成代码。

作为第一步，我们考虑每一对变量之间的样本相关系数。图 16-13 是利用 Minitab 的相关命令得到的相关矩阵。注意，Sales 和 Time 之间的样本相关系数是 0. 623，Sales 和 Poten 之间的样本相关系数是 0. 598，等等。

```
            Sales     Time     Poten    AdvExp    Share    Change   Accounts    Work
Time        0.623
Poten       0.598     0.454
AdvExp      0.596     0.249     0.174
Share       0.484     0.106    -0.21     0.264
Change      0.489     0.251     0.268    0.377     0.085
Accounts    0.754     0.758     0.479    0.200     0.403    0.327
Work       -0.117    -0.179    -0.259   -0.272     0.349   -0.288    -0.199
Rating      0.402     0.101     0.359    0.411    -0.024    0.549     0.229    -0.277
```

图 16-13　Cravens 数据的样本相关系数

我们观察一下自变量之间的样本相关系数，我们看到 Time 和 Accounts 之间的样本相关系数是 0. 758；因此，如果我们用 Accounts 作为一个自变量，那么 Time 将不会对模型增加更多的解释能力。回忆一下在第 15. 5 节中讨论多重共线性时的经验检验法则：对于任意两个自变量，如果样本相关系数的绝对值大于 0. 7，那么多重共线性可能对估计结果产生影响。因此，如果可能的话，我们应该避免 Time 和 Accounts 这两个自变量同时出现在一个回归模型中。Change 和 Rating 之间的样本相关系数是 0. 549，它也是比较高的，所以我们有理由要对这两个自变量做进一步考虑。

让我们观察一下 Sales 和每一个自变量之间的样本相关系数，很快我们就能知道，哪些自变量它们自己就是很好的预测变量。我们看到，Accounts 是 Sales 的一个最好的预测变量，因为它们之间的样本相关系数最高（0. 754）。回想一下，对于一个自变量的情形，样本相关系数的平方是判定系数。于是，Accounts 能解释 Sales 中的 $(0.754)^2 \times 100$，或者 56. 85% 的变异性。接下来最重要的自变量 Time、Poten 和 AdvExp，它们每一个和 Sales 之间的样本相关系数都近似等于 0. 6。

虽然存在潜在的多重共线性，我们还是考虑利用全部 8 个自变量，建立一个估计的回归方程。Minitab 计算机

软件包给出的计算结果如图 16-14 所示。8 个自变量的多元回归模型的修正判定系数是 88.3%。然而我们注意到，对于单个参数的 t 检验，在 $\alpha=0.05$ 的显著性水平下，当所有其他变量的影响已知时，仅有 Poten、AdvExp 和 Share 的 p-值是显著的。因此，我们或许倾向于研究仅仅利用这 3 个自变量而得到的模型。在图 16-15 中，Minitab 给出了利用这 3 个自变量得到的估计的回归方程的结果。我们看到，估计的回归方程的修正判定系数是 82.7%，尽管它不如利用全部 8 个自变量得到的估计的回归方程那样好，但这个修正判定系数也是非常高的。

```
The regression equation is
Sales = - 1508 + 2.01 Time + 0.0372 Poten + 0.151 AdvExp + 199 Share
        + 291 Change + 5.55 Accounts + 19.8 Work + 8 Rating

Predictor      Coef    SE Coef      T      p
Constant    -1507.8      778.6  -1.94  0.071
Time          2.010      1.931   1.04  0.313
Poten      0.037206   0.008202   4.54  0.000
AdvExp      0.15094    0.04711   3.21  0.006
Share        199.08      67.03   2.97  0.009
Change        290.9      186.8   1.56  0.139
Accounts      5.550      4.775   1.16  0.262
Work          19.79      33.68   0.59  0.565
Rating          8.2      128.5   0.06  0.950

S = 449.015   R-Sq = 92.2%   R-Sq(adj) = 88.3%

Analysis of Variance

SOURCE          DF        SS        MS      F      p
Regression       8  38153712   4769214  23.66  0.000
Residual Error  16   3225837    201615
Total           24  41379549
```

图 16-14 包括全部 8 个自变量模型的 Minitab 输出

```
The regression equation is
Sales = - 1604 + 0.0543 Poten + 0.167 AdvExp + 283 Share

Predictor      Coef    SE Coef      T      p
Constant    -1603.6      505.6  -3.17  0.005
Poten      0.054286   0.007474   7.26  0.000
AdvExp      0.16748    0.04427   3.78  0.001
Share        282.75      48.76   5.80  0.000

S = 545.515   R-Sq = 84.9%   R-Sq(adj) = 82.7%

Analysis of Variance

SOURCE          DF        SS        MS      F      p
Regression       3  35130228  11710076  39.35  0.000
Residual Error  21   6249321    297587
Total           24  41379549
```

图 16-15 包括 3 个自变量 Poten、AdvExp 和 Share 模型的 Minitab 输出

在可供使用的数据已知时，我们如何才能求得一个具有最佳效果的估计的回归方程呢？一个方法是计算所有可能的回归方程。也就是说，我们要建立 8 个单变量的估计的回归方程（每一个回归方程对应一个自变量），28 个两个变量的估计的回归方程（从 8 个变量中每次取 2 个的组合数），等等。对 Cravens 数据，总共有 255 个包含一个或一个以上自变量的不同的估计的回归方程可用于数据拟合。

现在，我们已经拥有一些更出色、更有效的计算机软件包，它能计算出所有可能的回归方程。但是，这样一来就要涉及大量的计算，并且要求模型设计者审查大量的计算机输出结果，显然大部分的输出结果都与不好的模型相联系。统计学家们更喜欢用系统的方法从全部自变量中选择一部分自变量，用这一部分自变量就能得到最优

的估计的回归方程。在下一节，我们将介绍几个比较通用的方法。

16.4　变量选择方法

在这一节，我们将讨论 4 种**变量选择方法**（variable selection procedure）：逐步回归，前向选择，后向消元和最佳子集回归。已知所有可能的自变量的数据集，我们可以使用这 4 种方法去确认哪些自变量能给出最佳的回归模型。前 3 种方法都是反复迭代的方法，在每一步只增加或者删除一个自变量，并对新的回归模型进行评价。这一迭代过程连续不断地进行下去，直到结束准则表明：用这种方法已不能找到一个更好的模型为止。最后一种方法（最佳子集回归）不是单变量一次性方法，这种方法要对包含不同的自变量子集的回归模型进行评价。[1]

在逐步回归，前向选择和后向消元方法的每一步中，是增加一个自变量到回归模型上，还是从回归模型中删除一个自变量，选择自变量的准则是我们在第 16.2 节中介绍的 F 统计量。例如，假设我们考虑增加 x_2 到包含 x_1 的模型上或者从包含 x_1 和 x_2 的模型中删除 x_2。为了检验增加或者删除的自变量 x_2 在统计上是否显著，我们提出的原假设和备择假设表示成如下形式。

$$H_0: \beta_2 = 0$$
$$H_a: \beta_2 \neq 0$$

在第 16.2 节，我们给出了下面的 F 统计量［见式（16-10）］：

$$F = \frac{\dfrac{\mathrm{SSE}(x_1) - \mathrm{SSE}(x_1, x_2)}{1}}{\dfrac{\mathrm{SSE}(x_1, x_2)}{n - p - 1}}$$

我们可以利用这个统计量作为一个准则，该准则能确定模型中 x_2 的存在是否会引起误差平方和有一个显著的减少。对应这个 F 统计量的 p-值也是一个准则，我们能利用这个准则来确定一个自变量是应该增加到回归模型上，还是应该从回归模型中删除。通常应用的拒绝法则是：如果 p-值 $\leqslant \alpha$，则拒绝 H_0。

16.4.1　逐步回归

逐步回归方法的每一步都是从确定已经在模型中的自变量是否应该被删除开始。为此，首先对已经在模型中的每一个自变量计算 F 统计量和对应的 p-值。为了确定一个自变量是否应该从模型中被删除的显著性水平 α，在 Minitab 输出中被表示为“*Alpha to remove*”。如果有自变量的 p-值大于“*Alpha to remove*”，则具有最大 p-值的自变量应该从模型中被删除，并且逐步回归方法开始新的一步。

如果没有自变量能从模型中被删除，那么逐步回归方法将试图使另一个自变量进入模型。为此，首先对没有在模型中的每一个自变量计算 F 统计量和对应的 p-值。为了确定一个自变量是否应该进入模型的显著性水平 α，在 Minitab 输出中被表示为“*Alpha to enter*”。如果有自变量的 p-值小于或等于“*Alpha to enter*”，则具有最小 p-值的自变量将进入模型。按照这种方式将逐步回归过程继续进行下去，直到没有一个自变量能从模型中被删除，或者没有一个自变量能被增加到模型上为止。

图 16-16 给出了对 Cravens 数据应用 Minitab 逐步回归程序得到的结果，“*Alpha to remove*”和“*Alpha to enter*”的值都取作 0.05。进行四步以后，逐步回归程序结束。根据 Minitab 逐步回归程序得到的估计的回归方程是

$$\hat{y} = -1\,441.93 + 9.2\mathrm{Accounts} + 0.175\mathrm{AdvExp} + 0.038\,2\mathrm{Poten} + 190\mathrm{Share}$$

在图 16-16 中我们还注意到，$s = \sqrt{\mathrm{MSE}}$ 从最佳单变量（仅含有一个自变量 Accounts）模型的 881，经过四步以后减少到 454。R-Sq 的值从 56.85% 增加到 90.04%，并且被推荐的估计的回归方程的 R-Sq（adj）的值是 88.05%。[2]

---

[1] 变量选择方法在建模的最初阶段对我们特别有帮助，但是这些方法不能代替分析工作者的经验和作出的判断。

[2] 在已知自变量个数的情况下，因为逐步回归方法并不考虑每一种可能的自变量组合子集，所以未必会选择具有最大 R-Sq 值的估计的回归方程。

综上所述，逐步回归方法的每一步，首先要考虑的是查看一下是否有哪个自变量能从当前的模型中被删除。如果没有一个变量能从模型中被删除，那么逐步回归方法要查看一下是否有哪个不在当前模型中的自变量能增加到模型里来。由于逐步回归方法的特性，一个自变量可能在回归过程的某一步进入模型，而在下一步这个自变量又可能从模型中被删除，但是在稍后的某一步它又可能重新进入模型。当没有自变量能从模型中被删除或者没有自变量能进入到模型里来时，逐步回归方法停止。

Alpha-to-Enter: 0.05　　Alpha-to-Remove: 0.05

Response is Sales on 8 predictors, with N = 25

Step	1	2	3	4
Constant	709.32	50.29	-327.24	-1441.93
Accounts	21.7	19.0	15.6	9.2
T-Value	5.50	6.41	5.19	3.22
P-Value	0.000	0.000	0.000	0.004
AdvExp		0.227	0.216	0.175
T-Value		4.50	4.77	4.74
P-Value		0.000	0.000	0.000
Poten			0.0219	0.0382
T-Value			2.53	4.79
P-Value			0.019	0.000
Share				190
T-Value				3.82
P-Value				0.001
S	881	650	583	454
R-Sq	56.85	77.51	82.77	90.04
R-Sq(adj)	54.97	75.47	80.31	88.05
Mallows Cp	67.6	27.2	18.4	5.4

图 16-16　Cravens 数据的 Minitab 逐步回归输出

16.4.2　前向选择

前向选择方法从模型中没有自变量开始。这一方法使用与逐步回归为了确定一个变量是否应该进入模型同样的程序来增加变量，并且一次只能增加一个变量。然而，一旦一个自变量进入到模型中，前向选择方法就不允许再将这个变量从模型中删除。当不在模型中的每一个自变量的 p- 值全都大于“*Alpha to enter*”时，则前向选择过程结束。

利用 Minitab 前向选择方法得到的估计的回归方程是

$$\hat{y} = -1\,441.93 + 9.2\text{Accounts} + 0.175\text{AdvExp} + 0.0382\text{Poten} + 190\text{Share}$$

于是，对于 Cravens 数据，当“*Alpha to enter*”的值取 0.05 时，前向选择方法与逐步回归方法得到完全相同的估计的回归方程。

16.4.3　后向消元

后向消元方法从包含所有自变量的模型开始。这一方法使用与逐步回归为了确定一个变量是否应该从模型中被删除同样的程序来删除变量，并且一次只能删除一个变量。然而，一旦一个自变量从模型中被删除，后向消元方法就不允许这个自变量在下一步再重新进入模型。当模型中自变量的 p- 值没有一个大于“*Alpha to remove*”时，则后向消元过程结束。

对于 Cravens 数据，当“*Alpha to remove*”的值取 0.05 时，利用 Minitab 后向消元方法得到的估计的回归方程是

$$\hat{y} = -1\,312 + 3.8\text{Time} + 0.0444\text{Poten} + 0.152\text{AdvExp} + 259\text{Share}$$

将使用后向消元方法得到的估计的回归方程与使用前向选择方法得到的估计的回归方程进行比较，我们看到在这两个方程中有 3 个自变量——AdvExp、Poten 和 Share 是共同的。然而，后向消元方法包含了自变量 Time，替代了前向选择方法的自变量 Accounts。

前向选择方法和后向消元方法是建模过程的两个极端情形。前向选择方法从模型中没有自变量开始，并且每次只增加一个变量。而后向消元方法从包含所有自变量的模型开始，并且每次只删除一个变量。两种方法可能得到同样的估计的回归方程，但是，正如我们在 Cravens 数据的例子中看到的那样，两种方法也可能得到不同的估计的回归方程。决定利用哪一个估计的回归方程将是留给我们讨论的话题。最后，必须运用分析人员的判断力。我们接着要讨论的最佳子集模型的建立过程，在作出最终决策之前就给出了补充的建模信息。⊖

⊖ 前向选择方法和后向消元方法可能会得出不同的模型。

16.4.4　最佳子集回归

逐步回归、前向选择和后向消元都是通过每次增加或者删除一个自变量来选择回归模型的方法。对于一组已知的变量，没有一种方法能保证我们将得到最佳的模型。对这些每次一个变量的单变量方法认真分析研究后，为我们选择一个好的回归模型提供了有益的启示。

一些计算机软件包使用被称为“最佳子集回归”的程序，对于一组给定的自变量，最佳子集回归程序能使软件包的使用者得到最佳回归模型。Minitab 就有这样的程序。图 16-17 是对 Cravens 数据集应用最佳子集回归程序得到的一部分计算机输出。[⊖]

在图 16-17 的输出中，识别出两个最佳的单变量估计的回归方程，两个最佳的两变量估计的回归方程，两个最佳的三变量估计的回归方程，等等。对于任一组预测变量，用于确定哪一个估计的回归方程是最佳的准则是判定系数（R-Sq）的数值。例如，因为 R-Sq = 56.8%，所以得到了仅含有 1 个自变量 Accounts 的最佳的估计的回归方程；因为 R-Sq = 77.5%，所以得到了含有 2 个自变量 AdvExp 和 Accounts 的最佳的估计的回归方程；因为 R-Sq = 84.9%，所以得到了含有 3 个自变量 Poten、AdvExp 和 Share 的最佳的估计的回归方程。对于 Cravens 数据，含有 6 个自变量：Time，Poten，AdvExp，Share，Change 和 Accounts 的模型的修正判定系数（R-Sq（Adj）= 89.4%）是最大的。可是，含有 4 个自变量（Poten，AdvExp，Share 和 Accounts）的最佳模型有一个差不多大小的修正判定系数（R-Sq（Adj）= 88.1%）。在所有其他条件相同的情形下，一个包含较少自变量的比较简单的模型通常会受到人们的喜爱。

```
                                          A
                                          c
                                    A   C c   R
                                  P d S h o   a
                                T o v h a u W t
                                i t E a n n o I
                                m e x r g t r n
Vars  R-Sq  R-Sq(adj)        S  e n p e e s K g

   1  56.8       55.0   881.09            X
   1  38.8       36.1   1049.3  X
   2  77.5       75.5   650.39      X     X
   2  74.6       72.3   691.11    X     X
   3  84.9       82.7   545.52    X X X
   3  82.8       80.3   582.64    X X     X
   4  90.0       88.1   453.84    X X X   X
   4  89.6       87.5   463.93  X X X X
   5  91.5       89.3   430.21  X X X X X
   5  91.2       88.9   436.75    X X X X X
   6  92.0       89.4   427.99  X X X X X X
   6  91.6       88.9   438.20    X X X X X X
   7  92.2       89.0   435.66  X X X X X X X
   7  92.0       88.8   440.29  X X X X X X   X
   8  92.2       88.3   449.02  X X X X X X X X
```

图 16-17　Minitab 最佳子集回归的一部分输出

16.4.5　作出最终的选择

到目前为止，对 Cravens 数据已经完成的分析为我们选择一个最终的模型做出了很好的准备，但是在作出最终选择前，我们还应进行更多的分析。正如我们在第 14 章和第 15 章中已经注意到的那样，我们应该对残差做出更细致的分析。在选择模型时，我们希望残差图看起来要近似于一条水平带。我们假设，残差满足这一要求，没有问题，并且希望使用最佳子集程序的结果帮助我们选择模型。

最佳子集方法已经向我们表明了最佳的 4 变量模型包含了 4 个自变量，即 Poten、AdvExp、Share 和 Accounts。这个最佳模型恰好也是利用逐步回归方法得到的包含 4 个自变量的模型。表 16-7 对于我们作出最终的选择是有帮助的。该表显示出上述 4 个自变量的一部分或者全部组成的各种可能的模型。

表 16-7　选择包含 ACCOUNTS、AdvExp、Poten 和 Share 的模型

模型	自变量	修正判定系数
1	Accounts	55.0
2	AdvExp，Accounts	75.5
3	Poten，Share	72.3
4	Poten，AdvExp，Accounts	80.3
5	Poten，AdvExp，Share	82.7
6	Poten，AdvExp，Share，Accounts	88.1

⊖ 完整的最佳子集输出还包括 *Mallows Cp* 统计量的值。更高级的教科书讨论这个统计量的应用。

从表16-7中我们看到，只包含2个自变量Accounts和AdvExp的模型是一个好的模型，它的修正判定系数是R-Sq(Adj) =75.5%，而包含了全部4个自变量模型的修正判定系数仅仅改善了12.6个百分点。例如，如果市场潜力（Poten）这个自变量的度量很困难，那么我们宁愿选择一个比较简单的两变量模型。然而，如果数据很容易取得，并且要求销售收入有一个较高的预测精度，模型的设计者显然应该更喜欢包含全部4个自变量的模型。

注释和评论

1. 逐步回归方法要求“*Alpha to remove*”的值大于或等于“*Alpha to enter*”的值。这一要求是为了防止同一个变量在同一步中先从模型中被删除，然后再次进入模型。
2. 我们在应用这一节介绍的任一种方法时，自变量的函数都可以用来生成新的自变量。例如，如果我们希望用x_1x_2表示模型中的交互作用，我们将利用自变量x_1和x_2的数据生成新的自变量$z=x_1x_2$的数据。
3. 每次只能增加或者删除一个变量的这些方法，没有一个方法能保证识别出最佳的回归模型。但是，这些方法都是得到好的模型的杰出方法—— 特别地，当存在轻微的多重共线性时。

练 习

应用

16. 一项研究提供了一些变量的数据，这些变量可能与制造业工人已经失业的周数有关。在研究中，应变量（Weeks）被定义为由于裁员，制造业工人已经失业的周数。在研究中用到的自变量如下所示。

Age	工人的年龄
Educ	受教育的年数
Married	虚拟变量。如果已婚，取值为1；否则，取值为0
Head	虚拟变量。如果是户主，取值为1；否则，取值为0
Tenure	工龄
Manager	虚拟变量。如果职业是从事管理工作，取值为1；否则，取值为0
Sales	虚拟变量。如果职业是从事销售工作，取值为1；否则，取值为0

可供我们使用的数据保存在本书所附光盘名为Layoffs的文件中。

a. 建立最佳的单变量的估计的回归方程。

b. 利用逐步回归方法建立最佳的估计的回归方程。“*Alpha to enter*”和“*Alpha to remove*”的值取0.05。

c. 利用前向选择方法建立最佳的估计的回归方程。“*Alpha to enter*”的值取0.05。

d. 利用后向消元方法建立最佳的估计的回归方程。“*Alpha to remove*”的值取0.05。

e. 利用最佳子集回归方法建立最佳的估计的回归方程。

18. Jeff Sagarin自1985年以来一直为《今日美国》杂志提供体育评级。在棒球比赛中，他预测的RPG（分/场）统计量考虑到全部球员的进攻统计，并且他声称：这是球员真实的进攻价值的最好的度量。下面给出了纽约洋基队（New York Yankees）的20名球员，在美国职业棒球大联盟2005年赛季中的RPG和各种各样的进攻统计数据（USA Today website，2006.3.3）。列标题定义如下：RPG，预测每场比赛得分；H，安打；2B，二垒打；3B，三垒打；HB，本垒打；RBI，击球得分；BB，四坏球后安全上垒（步行）；SO，三击不中出局；SB，偷垒；CS，封杀偷垒；OBP，上垒率；SLG，长打率，；AVG，安打率。

设RPG统计量为应变量。

a. 建立最佳的单变量的估计的回归方程。

b. 利用本节介绍的方法，建立一个能用来预测球员RPG的最佳的多元估计的回归方程。

球员	RPG	H	2B	3B	HB	RBI	BB	SO	SB	CS	OBP	SLG	AVG
D Jeter	6.51	202	25	5	19	70	77	117	14	5	0.389	0.450	0.309
H Matsui	6.32	192	45	3	23	116	63	78	2	2	0.367	0.496	0.305
A Rodriguez	9.06	194	29	1	48	130	91	139	21	6	0.421	0.610	0.321

（续）

球员	RPG	H	2B	3B	HB	RBI	BB	SO	SB	CS	OBP	SLG	AVG
G Sheffield	6.93	170	27	0	34	123	78	76	10	2	0.379	0.512	0.291
R Cano	5.01	155	34	4	14	62	16	68	1	3	0.320	0.458	0.297
B Williams	4.14	121	19	1	12	64	53	75	1	2	0.321	0.367	0.249
J Posada	5.36	124	23	0	19	71	66	94	1	0	0.352	0.430	0.262
J Giambi	9.11	113	14	0	32	87	108	109	0	0	0.440	0.535	0.271
T Womack	2.91	82	8	1	0	15	12	49	27	5	0.276	0.280	0.249
T Martinez	5.08	73	9	0	17	49	38	54	2	0	0.328	0.439	0.241
M Bellhorn	4.07	63	20	0	8	30	52	112	3	0	0.324	0.357	0.210
R Sierra	3.27	39	12	0	4	29	9	41	0	0	0.265	0.371	0.229
J Flaherty	1.83	21	5	0	2	11	6	26	0	0	0.206	0.252	0.165
B Crosby	3.48	27	0	1	1	6	4	14	4	1	0.304	0.327	0.276
M Lawton	5.15	6	0	0	2	4	7	8	1	0	0.263	0.250	0.125
R Sanchez	3.36	12	1	0	0	2	2	3	0	1	0.326	0.302	0.279
A Phillips	2.13	6	4	0	1	4	1	13	0	0	0.171	0.325	0.150
M Cabrera	1.19	4	0	0	0	0	0	2	0	0	0.211	0.211	0.211
R Johnson	3.44	4	2	0	0	0	1	4	0	0	0.300	0.333	0.222
F Escalona	5.31	4	1	0	0	2	1	4	0	0	0.375	0.357	0.286

16.5 实验设计的多元回归方法

在第15.7节中，我们已经讨论了虚拟变量在多元回归分析中的应用。在这一节中我们将说明，在多元回归方程中如何使用虚拟变量得到解决实验设计问题的另一种方法。在第13章中我们已经介绍了Chemitech公司的完全随机化设计，为了说明实验设计的多元回归方法，我们将对Chemitech公司应用这一方法。

表16-8 15名工人生产的过滤系统的数量

方法A	方法B	方法C
58	58	48
64	69	57
55	71	59
66	64	47
67	68	49

让我们回忆一下，Chemitech公司开发了一种新的城市供水过滤系统。新过滤系统的部件需要从几家供应商处购买，然后由Chemitech公司设在南加州哥伦比亚市的工厂装配这些部件。有三种不同的装配方法，分别称为方法A、方法B和方法C。Chemitech公司的管理人员想要确定：哪种装配方法能使每周生产的过滤系统数量最多。

抽取15名工人组成一个随机样本，并且对每一种装配方法随机地指派5名工人。每名工人装配的过滤系统的数量如表16-8所示。

三种装配方法所生产的过滤系统的样本均值如下所示。

装配方法	平均生产过滤系统的数量
A	62
B	66
C	52

虽然方法B似乎比其他方法能够带来更高的生产率，但我们的问题是：观测到的三个样本均值的差异是否足以使我们断定，对应于三种装配方法的总体均值是不同的。

为了用回归方法处理这一问题，我们从定义虚拟变量开始，这里的虚拟变量表示使用的是哪一种装配方法。因为Chemitech问题有三种装配方法或处理，所以我们需要定义两个虚拟变量。一般情况下，如果被研究的因素有k个不同的水平或处理，那么我们需要定义k-1个虚拟变量。对于Chemitech实验，我们定义的两个虚拟变量A

和 B 如表 16-9 所示。

我们可以利用虚拟变量将每周生产的过滤系统数量 y 与工人使用的装配方法联系起来。

$$E(y) = \text{每周生产的过滤系统的期望值} = \beta_0 + \beta_1 \text{A} + \beta_2 \text{B}$$

表 16-9 Chemitech 问题的虚拟变量

A	B	
1	0	观测值与方法 A 有关
0	1	观测值与方法 B 有关
0	0	观测值与方法 C 有关

于是，如果我们对使用方法 C 的一名工人每周装配的过滤系统的期望值感兴趣，我们的程序是指定虚拟变量 A 和 B 的数值，使得 A = B = 0。那么多元回归方程化简为

$$E(y) = \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 = \beta_0$$

我们能把 β_0 看作是使用方法 C 的一名工人每周装配的过滤系统的期望值。换句话说，β_0 是使用方法 C 的一名工人每周装配的过滤系统的平均值。

接下来，我们考虑使用其他装配方法的多元回归方程的形式。对于装配方法 A，虚拟变量的值是 A = 1 和 B = 0，于是有

$$E(y) = \beta_0 + \beta_1 \times 1 + \beta_2 \times 0 = \beta_0 + \beta_1$$

对于装配方法 B，我们令 A = 0 和 B = 1，于是有

$$E(y) = \beta_0 + \beta_1 \times 0 + \beta_2 \times 1 = \beta_0 + \beta_2$$

我们看到，$\beta_0 + \beta_1$ 是使用方法 A 的一名工人每周装配的过滤系统的平均值，$\beta_0 + \beta_2$ 是使用方法 B 的一名工人每周装配的过滤系统的平均值。

表 16-10 Chemitech 问题完全随机化设计的输入数据

A	B	y
1	0	58
1	0	64
1	0	55
1	0	66
1	0	67
0	1	58
0	1	69
0	1	71
0	1	64
0	1	68
0	0	48
0	0	57
0	0	59
0	0	47
0	0	49

现在我们希望估计系数 β_0、β_1 和 β_2，并且对于每一种方法，要求给出每周装配的过滤系统平均值的估计。表 16-10 是由 A、B 和 y 的 15 组观测值组成的样本数据。图 16-18 是对应的 Minitab 多元回归输出。我们看到，β_0、β_1 和 β_2 的估计值是 $b_0 = 52$、$b_1 = 10$ 和 $b_2 = 14$。于是，对于每一种方法，每周装配的过滤系统平均值的最佳估计如下所示：

装配方法	$E(y)$ 的估计值
A	$b_0 + b_1 = 52 + 10 = 62$
B	$b_0 = 52 + 14 = 66$
C	$b_0 = 52$

```
The regression equation is
y = 52.0 + 10.0 A + 14.0 B

Predictor      Coef   SE Coef       T      P
Constant     52.000     2.380   21.84  0.000
A            10.000     3.367    2.97  0.012
B            14.000     3.367    4.16  0.001

S = 5.32291    R-Sq = 60.5%    R-Sq(adj) = 53.9%

Analysis of Variance

SOURCE           DF       SS      MS      F      P
Regression        2   520.00  260.00   9.18  0.004
Residual Error   12   340.00   28.33
Total            14   860.00
```

图 16-18 Chemitech 问题的完全随机化设计的多元回归输出

注意，从回归分析中得到的每一种装配方法生产的过滤系统平均值的估计，与前面给出的样本均值是相同的。

现在让我们来看看，我们如何才能利用多元回归分析的输出，对这三种装配方法生产的过滤系统平均值之间的区别，进行 ANOVA 检验。首先我们注意到，如果在这些平均值之间没有区别，那么有

$$\text{方法 A 的 } E(y) - \text{方法 C 的 } E(y) = 0$$

$$\text{方法 B 的 } E(y) - \text{方法 C 的 } E(y) = 0$$

因为对于方法 C，$E(y)$ 等于 β_0，对于方法 A，$E(y)$ 等于 $\beta_0 + \beta_1$，所以方法 A 与方法 C 之间的差等于 $(\beta_0 + \beta_1) - \beta_0 = \beta_1$。又因为对于方法 B，$E(y)$ 等于 $\beta_0 + \beta_2$，所以方法 B 与方法 C 之间的差等于 $(\beta_0 + \beta_2) - \beta_0 = \beta_2$。我

们将有结论：如果$\beta_1=0$和$\beta_2=0$，那么三种装配方法没有差异。于是，检验三种装配方法均值差异的原假设可以叙述如下。

$$H_0:\beta_1=\beta_2=0$$

假设显著性水平是$\alpha=0.05$。我们还记得，为了检验这种类型的回归关系显著性的原假设，我们利用总体显著性的F检验。在图16-18的Minitab输出中，对应于$F=9.18$的p-值是0.004。因为p-值$=0.004<\alpha=0.05$，所以我们拒绝原假设H_0：$\beta_1=\beta_2=0$，并且得到结论：三种装配方法的均值是不同的。因为F检验表明，多元回归关系是显著的，所以可以进行t检验来确定个别参数β_1和β_2的显著性。在$\alpha=0.05$的显著性水平下，Minitab输出的p-值分别是0.012和0.001，这就表明我们可以拒绝H_0：$\beta_1=0$和H_0：$\beta_2=0$。于是，两个参数在统计上是显著的，从而，我们也可以得出这样的结论：方法A与方法C的均值是不同的，以及方法B与方法C的均值是不同的。

练习

方法

☆20. 考虑包括A、B、C和D四种处理的一个完全随机化设计。写出一个能用来分析该完全随机化设计数据的多元回归方程。对所有的变量给出定义。

22. 对于因素A有2水平，因素B有3水平的一个两因素设计，写出一个能用来分析该两因素设计数据的多元回归方程。对所有的变量给出定义。

应用

24. 对于4种不同的涂料，广告介绍它们的干燥时间是相同的。为了证实制造商的说法，对每一种涂料检测5个样本。对于每一个样本，记录了从涂上第一道涂料后经过充分干燥直到可以再涂第二道涂料的时间（单位：分钟）。得到的数据如下表所示。

涂料1	涂料2	涂料3	涂料4
128	144	133	150
137	133	143	142
135	142	137	135
124	146	136	140
141	130	131	153

a. 在$\alpha=0.05$的显著性水平下，检验4种不同涂料平均干燥时间是否存在显著差异。

b. 对于涂料2，你估计的平均干燥时间是多少？从计算机输出中得到的平均干燥时间是多少？

26. 在杂志上刊登广告的尺寸大小和广告的设计方案，可能会影响收到邮购商品广告目录的人数（单位：1 000人）。为了检验这种影响，一家经营邮购商品广告目录的公司设计了一种因子实验。考虑3种广告的设计方案和2种刊登广告的尺寸，得到的数据如下表所示。在$\alpha=0.05$的显著性水平下，检验因广告的设计方案、刊登广告的尺寸以及它们之间的交互作用而产生的显著影响。

		广告尺寸	
		小	大
方案	A	8	12
		12	8
	B	22	26
		14	30
	C	10	18
		18	14

16.6 自相关性和杜宾-瓦特森检验

通常，在工商管理和经济学的回归研究中，所利用的数据是按时间顺序采集的。我们用y_t表示y在时期t的值，而y_t的值又常常要依赖于y在以前时期的值。在这种情形下，我们说在数据中存在**自相关性**（autocorrelation），也叫**序列相关**（serial correlation）。如果y在t时期的值依赖于y在$t-1$时期的值，我们就说在数据中存在一阶自相关性。如果y在t时期的值依赖于y在$t-2$时期的值，我们就说在数据中存在二阶自相关性，等等。

回归模型的假定之一是模型的误差项是独立的。但是，当数据存在自相关性时，违背了这一假定。在一阶自

相关性情形中，t 时期的误差项 ε_t 将依赖于 $t-1$ 时期的误差项 ε_{t-1}。一阶自相关性的两种情形如图 16-19 所示。图 16-19a 是存在正自相关性的情形；图 16-19b 是存在负自相关性的情形。当存在正自相关性时，如果在一个时期有一个正的残差，我们期待着在下一个时期跟着也是一个正的残差；如果在一个时期有一个负的残差，我们期待着在下一个时期跟着也是一个负的残差，等等。当存在负自相关性时，如果在一个时期有一个正的残差，我们期待着在下一个时期跟着一个负的残差，然后又是一个正的残差，等等。

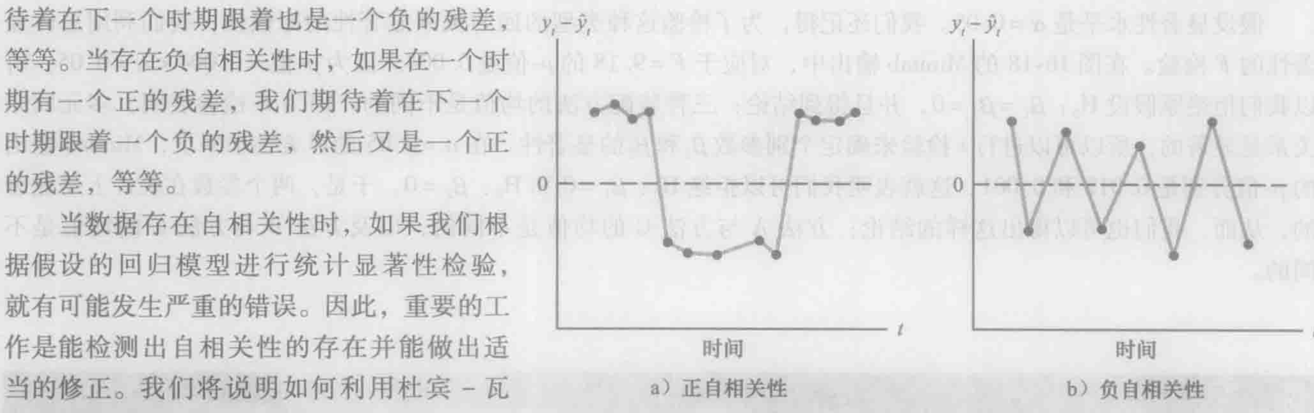


图 16-19　一阶自相关性的两个数据集

当数据存在自相关性时，如果我们根据假设的回归模型进行统计显著性检验，就有可能发生严重的错误。因此，重要的工作是能检测出自相关性的存在并能做出适当的修正。我们将说明如何利用杜宾 - 瓦特森统计量来检测一阶自相关性。

假设，误差项 ε 的值是不独立的，并且它们相互依赖的方式如下。

$$\varepsilon_t = \rho\varepsilon_{t-1} + z_t \tag{16-16}$$

这里，ρ 是一个绝对值小于 1 的参数，z_t 是一个平均值为 0、方差为 σ^2 的独立的正态分布的随机变量。从式（16-16）中我们看到，如果 $\rho=0$，那么误差项 ε_t 之间不相关，并且每一个误差项 ε_t 的平均值都是 0，方差都是 σ^2。在这种情形中，不存在自相关性，回归模型的假定被满足。如果 $\rho>0$，则存在正自相关性；如果 $\rho<0$，则存在负自相关性。这两种情形都违背了回归模型误差项的假定。

自相关性的**杜宾 - 瓦特森检验**（Durbin-Watson test）是利用残差来确定 $\rho=0$ 是否成立。为简化杜宾 - 瓦特森检验统计量的记号，我们用 $e_t=y_t-\hat{y}$ 表示第 i 个残差。杜宾 - 瓦特森检验统计量的计算公式如下。

杜宾 - 瓦特森检验统计量

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{16-17}$$

如果残差的相邻值彼此之间相距比较近（正自相关性），那么杜宾 - 瓦特森检验统计量的值将是比较小的。如果残差的相邻值彼此之间相距比较远（负自相关性），那么杜宾 - 瓦特森检验统计量的值将是比较大的。

杜宾 - 瓦特森检验统计量的取值范围介于 0 ~ 4，并且在 0 ~ 4 之间还有两个值表明了不存在自相关性的范围。杜宾和瓦特森给出了一个临界值表，这个表能用来确定，什么时候杜宾 - 瓦特森检验统计量的值表示存在自相关性。对于显著性水平 $\alpha=0.05$，表 16-11 给出了检验自相关性假设的下界和上界（d_L 和 d_U）；n 表示观测值的个数。检验的原假设始终是不存在自相关性。

$$H_0: \rho = 0$$

检验正自相关性的备择假设是

$$H_a: \rho > 0$$

检验负自相关性的备择假设是

$$H_a: \rho < 0$$

也允许进行双边检验。在这种情形中，备择假设是

$$H_a:\rho\neq 0$$

表 16-11　自相关性的杜宾－瓦特森检验的临界值

	α = 0.05 的 d_L 和 d_U 的显著点自变量的个数									
	1		2		3		4		5	
n^*	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

注：* 中间的 n 值，为线性插值。

说明：表中的值是自相关性的单侧杜宾－瓦特森检验的临界值。对于自相关性的双侧检验，显著性水平应增加 1 倍。

图 16-20 说明了如何利用表 16-11 中的 d_L 和 d_U 的值去检验自相关性。图 16-20a 说明了正自相关性的检验。如果 $d<d_L$，我们的结论是存在正自相关性。如果 $d_L\leqslant d\leqslant d_U$，我们的结论是杜宾－瓦特森检验不能确定是否存在正自相关性。如果 $d>d_U$，我们的结论是没有存在正自相关性的任何证据。

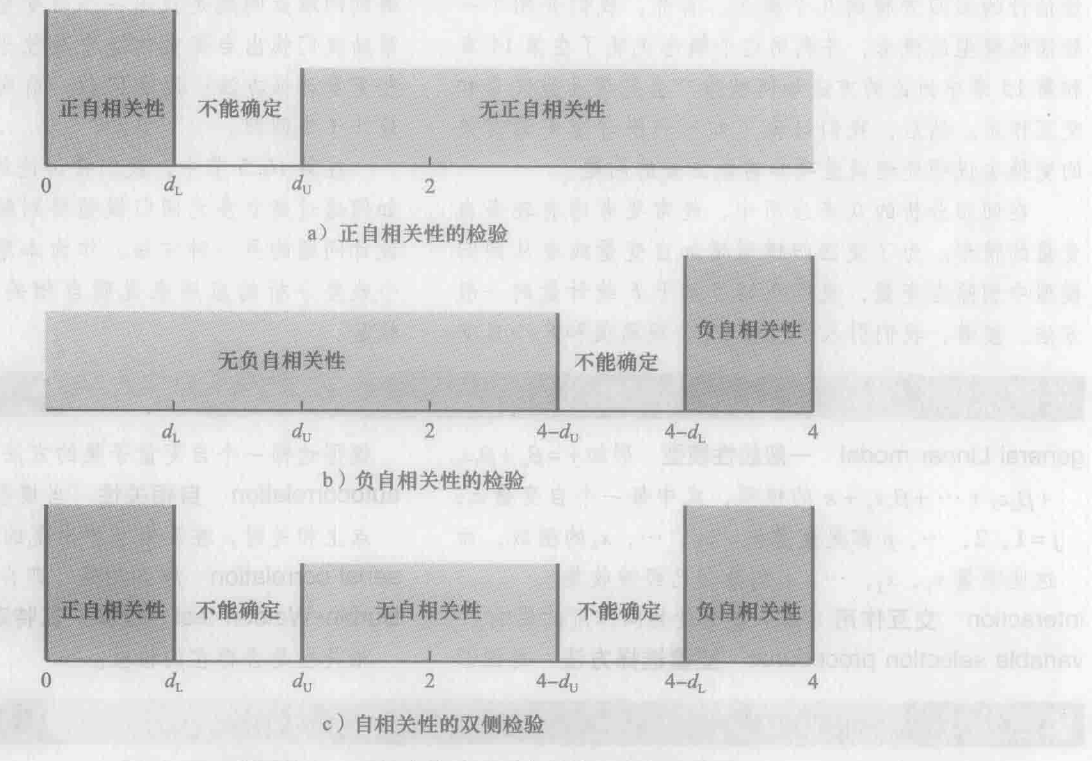


图 16-20　利用杜宾－瓦特森统计量对自相关性的假设检验

图 16-20b 说明了负自相关性的检验。如果 $d>4-d_L$，我们的结论是存在负自相关性。如果 $4-d_U\leqslant d\leqslant 4-d_L$，我们的结论是杜宾－瓦特森检验不能确定是否存在负自相关性。如果 $d<4-d_U$，我们的结论是没有存在负自相关性的任何证据。

图 16-20c 说明了双侧检验。如果 $d<d_L$ 或者 $d>4-d_L$，则拒绝 H_0，我们的结论是存在自相关性。如果 $d_L\leqslant d\leqslant d_U$ 或者 $4-d_U\leqslant d\leqslant 4-d_L$，我们的结论是杜宾－瓦特森检验不能确定是否存在自相关性。如果 $d_U<d<4-d_U$，我们的结论是没有存在自相关性的任何证据。

如果显著的自相关性被识别出来，我们应考虑假设的回归模型是否遗漏了一个或几个重要的自变量，而这些遗漏的自变量对应变量有显著的时序影响。如果没有这样的自变量被识别出来，那么可以在模型中引入一个度量观测次数的自变量（例如，对于第一次观测，这个变量的值可以为1，对于第二次观测，这个变量的值可以为2，等等），这样做将有助于消除或者减少自相关性。当这些消除或减少自相关性的尝试不起作用时，我们对应变量或者自变量进行适当的变换可能是有帮助的。有关这种变换的讨论能在更高级的回归分析教科书中找到。

注意，杜宾－瓦特森临界值表列出的最小样本容量为15。理由是当样本容量比较小时，检验通常是缺乏说服力的。事实上，许多统计学家认为，为了使检验能得到合理的结论，样本容量至少应该为50。

练习

应用

28. 参看表16-5中的Cravens数据集。在第16.3节中我们已经得到了含有变量Accounts、AdvExp、Poten和Share的估计的回归方程，该方程的修正判定系数为88.1%。在$\alpha=0.05$的显著性水平下，应用杜宾－瓦特森检验来确定是否存在正自相关性。

小结

在这一章中，我们讨论了模型设计者用来识别最佳估计的回归方程的几个概念。首先，我们介绍了一般线性模型的概念，并利用这个概念说明了在第14章和第15章中讨论的方法如何被推广去处理曲线关系和交互作用。然后，我们讨论了如何利用与应变量有关的变换去说明处理误差项非常数方差的问题。

在回归分析的众多应用中，经常要考虑有较多自变量的情形。为了使回归模型增加自变量或者从回归模型中删除自变量，我们介绍了基于F统计量的一般方法。接着，我们引入了包含25个观测值和8个自变量的一个大型问题。我们看到在处理一个大型问题时，遇到的难点问题是找出一个自变量的最佳子集。为了帮助我们找出自变量的这个最佳子集，我们讨论了一些变量选择方法：逐步回归、前向选择、后向消元和最佳子集回归。

在第16.5节中，我们将讨论的内容加以扩展，即如何通过建立多元回归模型得到解决方差分析和实验设计问题的另一种方法。作为本章的结束，我们用一个残差分析的应用来说明自相关性的杜宾－瓦特森检验。

关键术语

general Linear model　一般线性模型　形如$y=\beta_0+\beta_1 z_1+\beta_2 z_2+\cdots+\beta_p z_p+\varepsilon$的模型，式中每一个自变量$z_j$，$j=1, 2, \cdots, p$都是变量$x_1, x_2, \cdots, x_k$的函数，而这些变量$x_1, x_2, \cdots, x_k$的数据已经被收集。

interaction　交互作用　两个自变量共同作用的影响。

variable selection procedures　变量选择方法　对回归模型选择一个自变量子集的方法。

autocorrelation　自相关性　当模型误差项在连续时间点上相关时，在误差项中出现的相关性。

serial correlation　序列相关　即自相关性。

Durbin-Watson test　杜宾－瓦特森检验　确定一阶自相关性是否存在的检验。

重要公式

一般线性模型

$$y=\beta_0+\beta_1 z_1+\beta_2 z_2+\cdots+\beta_p z_p+\varepsilon \quad (16\text{-}1)$$

增加或删除$p-q$个变量的F检验统计量

$$F=\frac{\dfrac{\mathrm{SSE}(x_1,x_2,\cdots,x_q)-\mathrm{SSE}(x_1,x_2,\cdots,x_q,x_{q+1},\cdots,x_p)}{p-q}}{\dfrac{\mathrm{SSE}(x_1,x_2,\cdots,x_q,x_{q+1},\cdots,x_q)}{n-p-1}} \quad (16\text{-}13)$$

一阶自相关性

$$\varepsilon_t=\rho\varepsilon_{t-1}+z_t \quad (16\text{-}16)$$

杜宾－瓦特森统计量

$$d=\frac{\sum_{i=2}^{n}(e_i-e_{i-1})^2}{\sum_{i=1}^{n}e_i^2} \quad (16\text{-}17)$$

补充练习

30.《消费者报告》杂志检测了19款不同品牌和型号的公路自行车，健身自行车和舒适型自行车。公路自行车是为喜爱远程公路骑行旅行的人群设计的；健身自行车是为经常参加体育锻炼或每天需要上下班的人群设计的；舒适型自行车通常是为在平坦的路面上做休闲游的人群设计的。对19款被检测自行车的类型、重量（磅）和价格（美元）的统计资料如下所示（*Consumer Reports* website，2009年2月）

品牌和型号	类型	重量（磅）	价格（美元）
KleinRA^ve v	公路型	20	1 800
Giant OCR Composite 3	公路型	22	1 800
Giant OCR 1	公路型	22	1 000
Specialized Roubaix	公路型	21	1 300
Trek Pilot 2. 1	公路型	21	1 320
Cannondale Synapse 4	公路型	21	1 050
LeMond Poprad	公路型	22	1 350
Raleigh Cadent 1. 0	公路型	24	650
Giant FCR 3	健身型	23	630
Schwinn Super Sport GS	健身型	23	700
Fuji Absolute 2. 0	健身型	24	700
Jamis Coda Comp	健身型	26	830
Cannondale Road Warrior 400	健身型	25	700
Schwinn Sierra GS	舒适型	31	340
Mongoose Switchback SX	舒适型	32	280
GiantSedona DX	舒适型	32	360
Jamis Explorее 4. 0	舒适型	35	600
Diamondback Wildwood Deluxe	舒适型	34	350
Specialized Crossroads Sport	舒适型	31	330

a. 以重量为自变量，价格为应变量，对这些数据绘制散点图。简单线性回归模型是一个合适的模型吗？

b. 设x=重量，利用x和x^2作为两个自变量，建立一个估计的多元回归方程。

c. 定义两个自行车类型的虚拟变量如下：

Type-Fitness：如果自行车是健身型，取值为1；否则，取值为0。

Type-Comfort：如果自行车是舒适型，取值为1；否则，取值为0。

利用这两个虚拟变量建立一个估计的回归方程，使该方程在自行车类型已知时，能用来预测自行车的价格。请将得到的估计的回归方程与在（b）中得到的估计的回归方程进行比较。

d. 为了解释自行车类型与自行车重量之间可能存在的交互作用，建立一个新的估计的回归方程，使该方程在自行车的类型、自行车的重量以及自行车的重量与（c）中定义的每一个虚拟变量的交互作用已知时，能用来预测自行车的价格。估计的回归方程是自行车价格的最佳估计值吗？

32. 参阅第31题中的数据。考虑一个能用来预测Delay的模型，该模型仅以Industry为自变量。在$\alpha=0.01$的显著性水平下，检验数据中的正自相关性。

34. 曾经进行过的一项研究是调查顾客逛商店时的行为。顾客被分为以下三种类型：不爱逛商店的顾客、消遣性逛商店的顾客和特别喜爱逛商店的顾客。在这项研究中，对于每一位顾客，我们得到了他在商店里是否感到舒适惬意的一个度量。较高的得分表示顾客有较满意的舒适感。假设这项研究的数据如下表所示。在$\alpha=0.05$的显著性水平下，对这三种类型的顾客在逛商店时舒适程度的差异进行检验。

不爱逛商店的顾客	消遣性逛商店的顾客	特别喜爱逛商店的顾客
4	5	5
5	6	7
6	5	5
3	4	7
3	7	4
4	4	6
5	6	5
4	5	7

案例16-1 职业高尔夫球协会巡回赛的统计分析

美国职业高尔夫球协会（PGA）保存了高尔夫球员参加PGA巡回赛的成绩和奖金的统计资料。在PGA巡回赛的赛事中，总奖金前125位的高尔夫球员可以获得下个赛季的免资格赛的特权。制作前125位高尔夫球员的奖

金排行榜是重要的，因为一名拥有“免资格赛”特权的高尔夫球员能直接参加下个赛季PGA巡回赛的全部比赛。

Scoringaverage 通常被认为是在PGA巡回赛中获得胜利的最重要的统计指标。为了研究平均击球杆数与诸如发球距离、发球准确度、标准杆上果岭、沙坑救球和每轮比赛推杆入洞的平均次数等变量之间的关系，在2008年PGA巡回赛的赛事中，总奖金前125位高尔夫球员的年终成绩的统计资料，保存在本书所附光盘名为PGATour的文件中（PGA Tour website，2009年）。数据集的每一行对应着参加PGA巡回赛的一位高尔夫球员，并且数据已按总奖金排序。数据集中的变量描述如下。

Money：参加PGA巡回赛赛事的总奖金。

Scoring Average：每轮比赛的平均击球杆数。

DrDist（Driving Distance，发球距离）：DrDist是每次发球实测的平均码数。在PGA巡回赛中，发球距离是在每轮比赛的两个球洞上测量的。谨慎选择两个相反方向的球洞，以抵消风力的影响。发球距离测量是待球处于静止状态后进行，而不管球是否在球道上。

DrAccu（Driving Accuracy，发球准确度）：高尔夫球员在发球处将球击上球道次数的比率（而不管是何种球杆）。发球准确度是对每个球洞测量的，但不包括标准杆是3杆的情形。

GIR（Greens in Regulation，标准杆上果岭）：高尔夫球员能够标准杆上果岭次数的比率。如果按标准杆上果岭（GIR）规定的杆数击球后，高尔夫球的任一部分触及果岭的推杆区域，则认为是标准杆上果岭。标准杆上果岭规定的杆数被定义为比标准杆少2杆上果岭（若标准杆为3杆洞，则第1杆上果岭；若标准杆为4杆洞，则第2杆上果岭；若标准杆为5杆洞，则第3杆上果岭）。换句话说，如果高尔夫球员比标准杆少2杆上果岭的推杆区域，就认为是标准杆上果岭。

Sand Saves：一旦高尔夫球落到靠近果岭的沙坑里，高尔夫球员能克服“地面的高低起伏”将球救出的比例（不考虑得分）。克服“地面的高低起伏”将球救出，表示球员用2杆或少于2杆将高尔夫球从靠近果岭的沙坑击入球洞。

PPR（Putts Per Round）：每轮比赛推杆入球洞的平均次数。

Scrambling：高尔夫球员没能做到标准杆上果岭，但还是取得标准杆或较好成绩的次数的比率。

管理报告

假设你被PGA巡回赛的专员聘请，准备在PGA巡回赛的年度会议上做数据分析的报告。专员问：是否有可能利用这些数据来确定高尔夫球员的成绩，也就是确定高尔夫球员平均击球杆数的最佳预测值。利用在本章和第15章介绍的方法分析数据。为PGA巡回赛的专员编写一份报告，这份报告总结了你的分析，包括关键的统计结果、结论和建议，附录中有对你编写报告适用的一些技术资料。

案例16-2 产自意大利皮埃蒙特地区的葡萄酒评级

Wine Spectator 杂志刊登的文章和评论涉及酿酒工业的各个方面，包括来自世界各地葡萄酒的评级。在最近一期，该杂志使用百分制，对产自意大利皮埃蒙特地区的475瓶葡萄酒进行了检测和评级（*Wine Spectator*，2011年4月30日）。下表是 *Wine Spectator* 杂志对送来检测的每一瓶酒，被评为经典的、优秀的、非常好、好、平庸、还是不推荐的理由。

分数	等级
95~100	经典的：极好的葡萄酒
90~94	优秀的：出众的特性和风格的葡萄酒
85~89	非常好：有着特殊气质的葡萄酒
80~84	好：制作精良的、可信赖的葡萄酒
75~79	平庸：可以饮用的、可能有轻微瑕疵的葡萄酒
75以下	不推荐

对大多数消费者而言，一个关键问题是：一瓶酒付的钱越多，这瓶酒就越好。为了调查这个问题，我们从产自意大利皮埃蒙特地区送到 *Wine Spectator* 杂志检测的475瓶葡萄酒中，选取100瓶组成一个随机样本。样本中每一瓶酒的价格（美元）、*Wine Spectator* 分数和等级的数据，保存在本书所附光盘名为 WineRatings 的文件中。

管理报告

1. 绘制一张表，按经典的、优秀的、非常好、好、平庸和不推荐将样本中的100瓶葡萄酒分类，并且表明每类葡萄酒的数量和平均价格。在葡萄酒的价格和 *Wine Spectator* 等级之间是否显现出任何关系？你关于数据的最初结论，在其他方面是否有引人注目之处？

2. 以价格为横轴，以 *Wine Spectator* 为纵轴，绘制散点图。在葡萄酒的价格和 *Wine Spectator* 等级之间是否显现出线性关系？

3. 利用线性回归，建立一个估计的回归方程，使该方程在葡萄酒的价格已知时，能用来预测 *Wine Spectator* 的等级。

4. 利用一个二阶模型，建立一个估计的回归方程，使该方程在葡萄酒的价格已知时，能用来预测 *Wine Spectator* 的等级。

5. 将线性拟合模型与二阶拟合模型进行比较。

6. 作为二阶拟合模型的替代，利用葡萄酒价格的自然对数做自变量，建立一个拟合模型。将这个拟合模型与二阶拟合模型进行比较。

7. 基于你的分析，你是否会认为：一瓶酒付的钱越多，这瓶酒就越好？

8. 假设，你打算为买一瓶酒最多花30美元。在本案例中，你是否会花比你的上限低得多的价格买一瓶酒？

第 17 章

时间序列分析及预测

CHAPTER 17

实践中的统计

内华达职业健康诊所[1]

内华达州，斯帕克斯

内华达职业健康诊所是位于内华达州斯帕克斯市的一家私人医疗诊所。该诊所已有20多年的历史，专门从事工业医疗，并且一直处于快速增长中。在26个月里，它的月营业额从57 000美元一直增长到超过300 000美元，直到主要大楼被大火烧毁。

诊所的保险单包括实物财产和设备，还包括由于正常商业经营被打断而遭受的经济损失。确定实物财产和设备在火灾中的损失额，在受理保险索赔中并不是一件很困难的事情。但是，确定该诊所在7个月的重建期间内的经营损失则是一件相当复杂的事情，它涉及业主和保险公司之间的谈判。没有预先制定的规则可以用来计算，假如没有发生火灾，诊所的营业额"将会有多少"。为了估计收入损失，诊所利用一种预测方法来测算在7个月的停业期间可能实现的营业增长。火灾前营业额的历史数据将会成为预测模型的基础，该模型具有线性趋势和季节成分，这些成分将在本章中加以讨论。这个预测模型使诊所能准确估计所蒙受的损失，并且最终被保险公司所接受。

本章的目的是介绍时间序列分析及预测。假设我们要提供公司的一种产品在来年的季度销售预测，生产计划、原材料的采购、库存策略和销售定额都将受到我们提供的季度预测的影响。因此，较差的预测将导致较差的计划并增加公司的生产成本。我们应该如何着手提供季度的销售预测呢？对经济状况好的判断、直觉和意识，可以给我们一个未来可能发生态势的粗略想法或"感觉"，但将这些感觉转化为能用来预测来年销售的数字是很困难的。

预测方法可分为定量和定性两种方法。定性方法通常利用专家判断来进行预测。当被预测变量的历史数据不适合或者难以获得时，可以使用这些方法。当以下条件同时满足时，可以使用定量预测方法：①被预测变量过去的信息可用；②这些信息可以被量化；③过去的模式将会持续到未来的假定合理。在这种情形下，可以使用时间序列法或因果法来进行预测。本章我们将专门关注定量预测方法[2]。

如果历史数据局限于被预测变量的过去值，这种预测方法被称为时间序列方法，历史数据被称为时间序列。时间序列分析的目的是在历史资料或时间序列中发现规律性的模式，然后将这个模式外推到未来。这种预测仅仅依赖于变量的过去值和（或）过去的预测误差。

因果预测方法的依据是假定我们正预测的变量与其他一个或几个变量存在一个因果关系。在第14章、第15章和第16章回归分析的讨论中，我们演示了如何用一个或多个自变量来预测单个应变量的值。将回归分析作为预测工具，我们可以将想要预测的时间序列的值作为应变量。因此，如果我们能确定相关的自变量或解释变量的一个合适的集合，或许就能够建立用于推测或预测时间序列的估计的回归方程。例如，许多产品的销售量受广告支出的影响，因此，可用回归分析建立一个方程，来显示销售量和广告支出的关系。一旦确定了下一个时期的广告预算，我们将这个数值代入方程，可以得到下一个时期销售量的推测值或预测值。注意，如果使用时间序列方法进行预测，则不需要考虑广告支出，也就是说，时间序列方法仅仅依赖于过去的销售资料。

通过将时间视为自变量，时间序列视为应变量，回归分析也可以用于时间序列方法。为了区分回归分析在两种情形下的应用，我们用术语截面回归和时间序列回归表述。于是，时间序列回归是指自变量是时间的回归分析应用。本章由于我们关注时间序列方法，因此，我们将不讨论因果预测方法的回归分析，这种方法已在前面关于预测的教材中讨论过。

[1] 作者感谢为"实践中的统计"提供了本案例的内华达职业健康诊所业务主管 Bard Betz 和行政执行助理 Curtis Brauer。

[2] 预测仅仅是对将来可能发生情况的一种简单的判断。管理者必须学会接受这样一种事实：无论使用什么样的预测方法，他们都不可能获得理想的预测结果。

17.1 时间序列的模式

时间序列（time series）是一个变量在连续时点或连续时期上测量的观测值的序列。测量可以每隔一小时、一天、一周、一个月、一年，或在任何其他规定的时间间隔上[⊖]进行。数据的模式是了解时间序列过去行为的重要因素。如果这种行为预计可以持续到未来，我们可以用过去的模式来指导我们选择合适的预测方法。

为了确定数据中的基本模式，有效的第一步是绘制**时间序列图**（time series plot）。时间序列图是时间和时间序列变量之间关系的图形表述：时间位于横轴，时间序列值位于纵轴。当观察时间序列图时，让我们回顾一下能被确认的数据模式的一些常见类型。

17.1.1 水平模式

当数据围绕着一个不变的均值上下波动时，则存在**水平模式**（horizontal pattern）。为了说明时间序列拥有水平模式，考虑表17-1中的12周的数据。这些数据是佛蒙特州本宁顿的一个汽油批发商销售汽油的数量。时间序列的平均数或均值为每周19.25或19 250加仑。图17-1是这些数据的时间序列图。注意，这些数据围绕着样本均值19 250加仑上下波动。尽管呈现随机波动，但我们仍然说这些数据有水平模式。

表17-1 汽油销售量的数据序列

周	销售量（1 000加仑）	周	销售量（1 000加仑）	周	销售量（1 000加仑）
1	17	5	18	9	22
2	21	6	16	10	20
3	19	7	20	11	15
4	23	8	18	12	22

术语**平稳时间序列**（stationary time series）[⊜]是指统计性质是与时间独立的时间序列，特别地，其含义为

1. 过程产生的数据有一个不变的均值；
2. 时间序列的变异性随时间的推移保持不变。

平稳时间序列的时间序列图总是显示水平模式。但仅仅观测到水平模式就得出时间序列是平稳的结论，并没有足够的证据。在关于预测的更高级的教材中讨论了确定时间序列是否平稳的步骤，并给出将一个不平稳的时间按序列转化为平稳序列的方法。

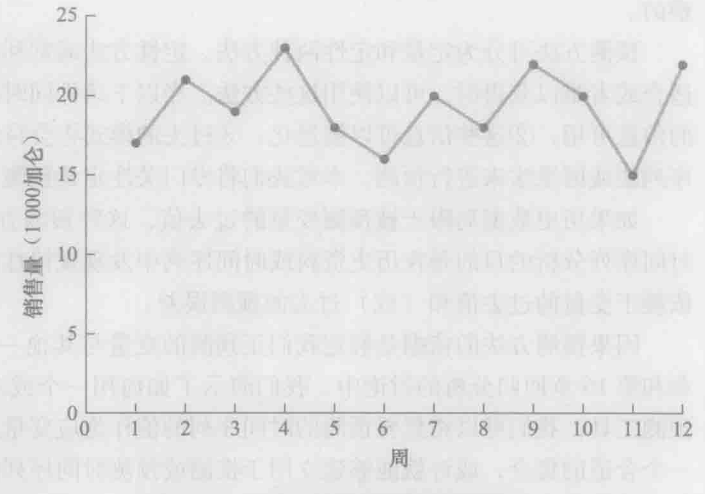


图17-1 汽油销售量的时间序列图

经验环境的改变常常导致有水平模式的时间序列移动到一个新的水平。例如，假设汽油批发商与佛蒙特州警方签署了一份合同，给佛蒙特州南部的警车加油。由于这个新合同，批发商希望看到从第13周开始周销售量有大幅度的增加。表17-2中给出了汽油销售量的原始时间序列以及签署了新合同后10周的销售量数据。图17-2是相应的时间序列图。注意从第13周开始时间序列水平增加了，时间序列水平的改变使得选择一个合适的预测方法变得更为困难。选择能较好地适应时间序列水平改变的预测方法在许多实际应用中是一件重要的值得考虑的事情。

⊖ 我们对时间序列的讨论仅限于时间序列的观测值是来自等间隔的情形，时间序列的观测值来自不等间隔的情形超出本书的讨论范围。

⊜ 平稳的正式定义见 G. E. P. Box，G. M. Jenkins 和 G. C. Reinsell，Time Series Analysis：Forecasting and Control，3rd ed. Englewood Cliffs，NJ：Prentice Hall，1994，p. 23.

表 17-2　获得佛蒙特州警方合同后汽油销售量的数据序列

周	销售量（1 000 加仑）	周	销售量（1 000 加仑）	周	销售量（1 000 加仑）
1	17	9	22	17	28
2	21	10	20	18	32
3	19	11	15	19	30
4	23	12	22	20	29
5	18	13	31	21	34
6	16	14	34	22	33
7	20	15	31		
8	18	16	33		

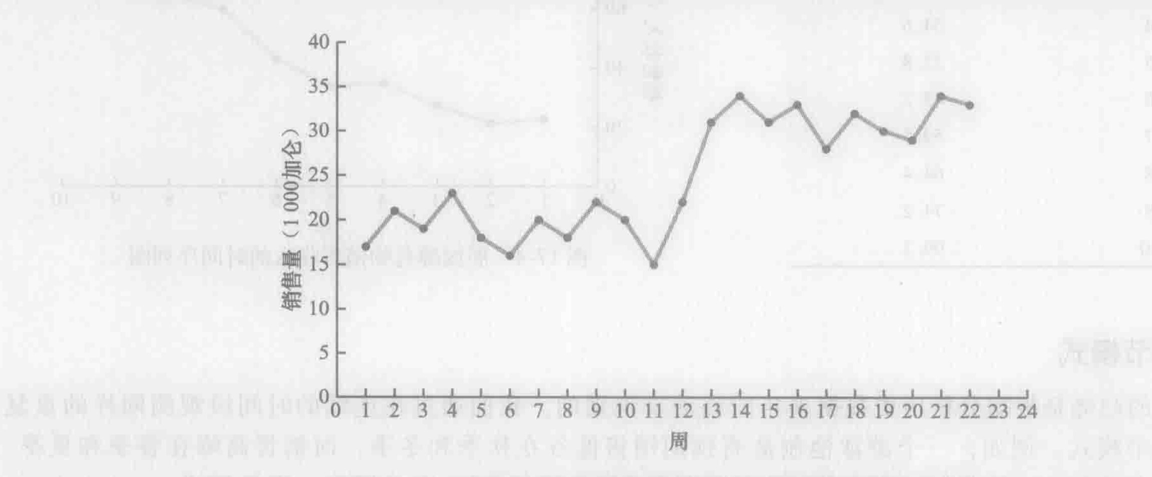


图 17-2　获得佛蒙特州警方合同后汽油销售量的时间序列图

17.1.2　趋势模式

尽管时间序列的数据通常呈现随机起伏的状态，但在一段较长的时间内，它仍然呈现出逐步的改变或移动到相对较高或较低的值。如果时间序列图显示出这类形态特征，我们则称存在**趋势模式**（trend pattern）。趋势通常是长期因素影响的结果，例如人口总数的增加或减少、人口总体统计特征、工艺和顾客偏爱的变化等。

为了说明拥有趋势模式的时间序列，我们考虑某自行车制造厂过去 10 年自行车销售量的时间序列，如表17-3 和图 17-3 所示。注意第 1 年自行车的销售量为 21 600 辆，第 2 年自行车的销售量为 22 900 辆，…，第 10 年（最近一年）自行车的销售量为 31 400 辆。时间序列图直观显示出，在过去 10 年中销售量有上下波动，但时间序列似乎总体是增长的或有向上的趋势。

表 17-3　自行车销售量的时间序列

年	销售量（1 000 辆）
1	21.6
2	22.9
3	25.5
4	21.9
5	23.9
6	27.5
7	31.5
8	29.7
9	28.6
10	31.4

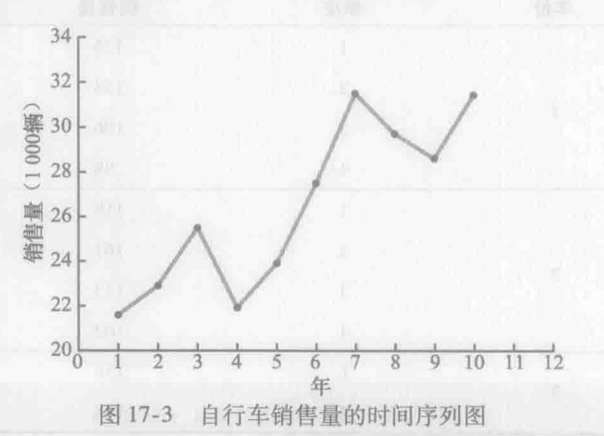


图 17-3　自行车销售量的时间序列图

自行车销售量时间序列的趋势呈现线性且随时间推移而增长，但是有时趋势能用其他类型的模式更好地描述。例如，表 17-4 给出了自从 10 年前公司的胆固醇药物获得 FDA 批准后的销售收入数据，图 17-4 是相应的时间序列图。时间序列以非线性的方式增长，即每年收入的变化率不是以一个固定的数值来增加。事实上，收入呈现指数形式的增长。当环比变化率从一个时期到下一个时期相对稳定时，这样的指数关系是适合的。

表 17-4 胆固醇药物销售收入的时间序列

年	销售收入（100 万美元）
1	23.1
2	21.3
3	27.4
4	34.6
5	33.8
6	43.2
7	59.5
8	64.4
9	74.2
10	99.3

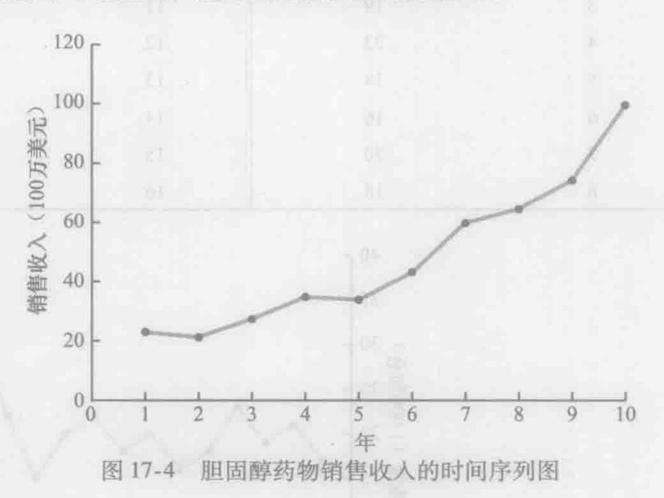


图 17-4 胆固醇药物销售收入的时间序列图

17.1.3 季节模式

时间序列的趋势是根据分析历史数据多年的移动来识别的。我们通过在连续的时间段观测同样的重复模式来确认季节模式。例如，一个游泳池制造商预期销售低谷在秋季和冬季，而销售高峰在春季和夏季。但是，铲雪设备和御寒衣物的制造商对他们产品销量的预期正好相反。毫无疑问，**季节模式**（seasonal pattern）是指，在超过一年的周期内，由于受季节的影响，时间序列呈现重复模式。尽管我们通常认为时间序列的季节变动是在一年内出现的，但在小于一年的时期内时间序列数据也可能呈现季节模式。例如，每天的交通流量数据显示在一天内的“季节”情况，高峰流量在上下班期间，中等流量在白天和傍晚，小流量是从午夜到清晨。

作为季节模式的一个例子，我们考虑过去 5 年一家商店的雨伞销售量。表 17-5 是时间序列，图 17-5 是相应的时间序列图。时间序列图没有显示销售量有长期趋势。事实上，如果不仔细观测数据，你可能会得出数据有水平模式的结论。但是，仔细检查时间序列图，你会发现数据有一个规律，即第一季度和第三季度销售量中等，第二季度销售量最高，第四季度销售量最低。因此，我们得出结论：存在季度的季节模式。

表 17-5 雨伞销售量的时间序列

年份	季度	销售量	年份	季度	销售量
1	1	125	3	3	113
	2	153		4	80
	3	106	4	1	109
	4	88		2	137
2	1	118		3	125
	2	161		4	109
	3	133	5	1	130
	4	102		2	165
3	1	138		3	128
	2	144		4	96

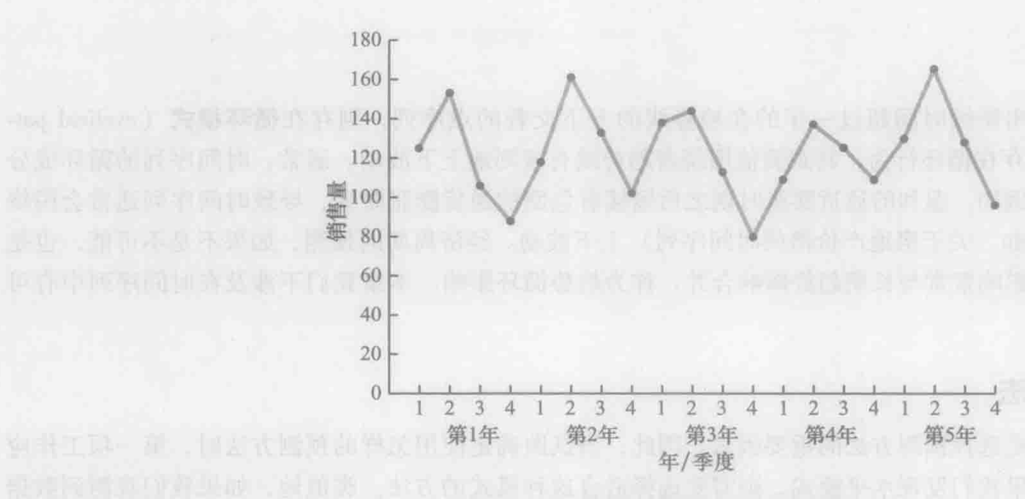


图 17-5　雨伞销售量的时间序列图

17.1.4　趋势与季节模式

有时时间序列同时包含趋势模式和季节模式。例如，表 17-6 中的数据是某制造商过去 4 年的电视机的销售量，图 17-6 是相应的时间序列图。显然，有增长趋势，但图 17-6 也显示出每年二季度销售量最低，第三季度和第四季度销售量上涨。于是，我们得出电视机销售量也存在季节模式。在这种情形下，我们需要用能同时处理趋势和季节性的预测方法。

表 17-6　电视机销售量的季度时间序列

年份	季度	销售量（1 000 台）	年份	季度	销售量（1 000 台）
1	1	4.8	3	1	6.0
	2	4.1		2	5.6
	3	6.0		3	7.5
	4	6.5		4	7.8
2	1	5.8	4	1	6.3
	2	5.2		2	5.9
	3	6.8		3	8.0
	4	7.4		4	8.4

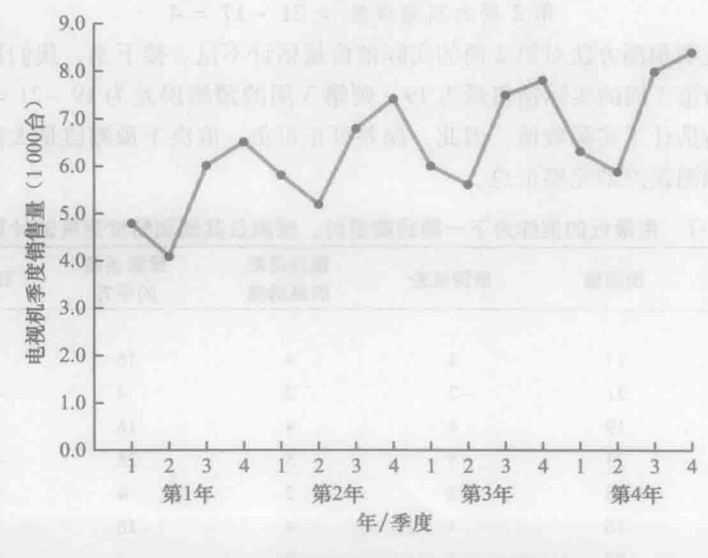


图 17-6　电视机季度销售量的时间序列图

17.1.5 循环模式

如果时间序列图显示出持续时间超过一年的在趋势线的上下交替的点序列，则存在**循环模式**（cyclical pattern）。许多经济时间序列存在循环行为，其观测值围绕着趋势线有规则地上下波动。通常，时间序列的循环成分归因于多年的经济周期。例如，温和的通货膨胀时期之后紧接着急剧的通货膨胀时期，导致时间序列通常会围绕着一条增长的趋势线（例如，关于房地产价格的时间序列）上下波动。经济周期的预测，如果不是不可能，也是极为困难的，因此，循环影响常常与长期趋势影响合并，称为趋势循环影响。本章我们不涉及在时间序列中有可能存在的循环影响。

17.1.6 选择预测方法

时间序列的基本模式是选择预测方法的重要因素。因此，当试图确定使用怎样的预测方法时，第一项工作应该是绘制时间序列图。如果我们发现水平模式，则需要选择适合这种模式的方法。类似地，如果我们观测到数据的趋势，则需要使用能有效地处理趋势的预测方法。下面两节将说明基本模式的水平模式（没有趋势或季节影响存在）时所使用的方法。然后，我们考虑在数据中存在趋势和（或）季节时的适合方法。

17.2 预测精度

本节首先我们对表17-1汽油销售量的时间序列用所有预测方法中最简单的一种方法进行预测：即用最近一周的销售量作为下一周的预测值。例如，批发商第1周售出17千加仑汽油；用这个值作为第2周的预测值。接着，我们用第2周的实际销售量21作为第3周的预测值，依此类推。用这种方法得到的预测值列在表17-7的预测值一列中。由于其简单，故该方法常常被称为朴素预测法。

用朴素预测法得到的预测精度如何？为了回答这个问题，我们介绍几种测量预测精度的方法。这些方法用于确定一种预测方法如何能很好地再现已经得到的时间序列数据。对于已知数据，通过选择最佳精度的方法，我们希望增加获得未来时期更好预测的可能性。

与测量预测精度相关的重要概念是**预测误差**（forecast error），其定义为

$$\text{预测误差} = \text{实际值} - \text{预测值}$$

例如，由于经销商第2周实际售出21千加仑汽油，而用第1周销售量得到的预测值是17千加仑，因此第2周的预测误差为

$$\text{第2周的预测误差} = 21 - 17 = 4$$

预测误差为正的事实表明预测方法对第2周的实际销售量估计不足。接下来，我们用第2周的实际销售量21作为第3周的预测值，因为第3周的实际销售量为19，则第3周的预测误差为 $19 - 21 = -2$。这时，负的预测误差表明第3周的预测值过高估计了实际数值。因此，误差可正可负，取决于预测值是太低还是太高。表17-7中预测误差一列是朴素预测法预测误差的完整汇总。

表17-7 用最近的值作为下一期预测值时，预测及其预测精度测度的计算表

周	时间序列值	预测值	预测误差	预测误差的绝对值	预测误差的平方	百分数误差	百分数误差的绝对值
1	17						
2	21	17	4	4	16	19.05	19.05
3	19	21	-2	2	4	-10.53	15.53
4	23	19	4	4	16	17.39	17.39
5	18	23	-5	5	25	-27.78	27.78
6	16	18	-2	2	4	-12.50	12.50
7	20	16	4	4	16	20.00	20.00
8	18	20	-2	2	4	-11.11	11.11

（续）

周	时间序列值	预测值	预测误差	预测误差的绝对值	预测误差的平方	百分数误差	百分数误差的绝对值
9	22	18	4	4	16	18.18	18.18
10	20	22	−2	2	4	−10.00	10.00
11	15	20	−5	5	25	−33.32	33.32
12	22	15	7	7	49	31.82	31.82
		合计	5	41	179	1.19	211.69

一个简单的预测精度的测度是预测误差[㊀]的均值或平均数。由表 17-7 可知，汽油销售量时间序列预测误差之和为 5，因此，预测误差的均值或平均数为 5/11 =0.45。注意，尽管汽油时间序列有 12 项，但计算平均误差时，我们用预测误差之和除以 11，这是因为只有 11 个预测误差。由于平均预测误差为正，这种方法有些估计不足；换句话说，观测值有大于预测值的倾向。由于正的和负的预测误差相互抵消，平均误差应该很小；因此，平均误差不是预测精度的常用测度。

平均绝对误差（mean absolute error）是避免正负预测误差相互抵消的一种预测精度的测量，记为 MAE。正如其名，MAE 是预测误差绝对值的平均数。由表 17-7 可知，预测误差绝对值之和为 41，因此

$$\text{MSE} = \text{预测误差绝对值的平均数} = \frac{41}{11} = 3.73$$

避免正负预测误差相互抵消的另一种测量方法是计算预测误差平方的平均数。这种预测精度的测度称为**均方误差**（mean squared error），记为 MSE[㊁]。由表 17-7 可知，误差平方之和为 179，因此

$$\text{MSE} = \text{预测误差平方和的平均数} = \frac{179}{11} = 16.27$$

MAE 和 MSE 的大小依赖于数据的测量尺度，因此，对不同时间间隔进行比较是很困难的。例如对预测月汽油销售量的方法和预测周销售量的方法进行比较，或对不同时间序列的比较都是困难的。为了对这些情况进行比较，我们需要考虑相对的或百分数误差的测度。**平均绝对百分数误差**（mean absolute percentage error），记为 MAPE，就是这样一种测度。为了计算 MAPE，首先我们必须计算每一个预测的百分数误差。例如，对应于第 2 周预测值 17 的百分数误差，是第 2 周的预测误差除以第 2 周的实际值再乘以 100。第 2 周的百分数误差计算如下：

$$\text{第 2 周的百分数误差} = \frac{4}{21} \times 100 = 19.05\%$$

于是，第 2 周的预测误差是第 2 周观测值的 19.05%。百分数误差的完整汇总列在表 17-7 百分数误差一列中。我们还给出了百分数误差的绝对值。

由表 17-7 可知，百分数误差的绝对值之和为 211.69，因此

$$\text{MAPE} = \text{百分数预测误差的绝对值的平均数} = \frac{211.69}{11} = 19.24\%$$

总之，用朴素（最近一期观测值）预测法，我们得到如下预测精度的测度：

$$\text{MAE} = 3.73$$
$$\text{MSE} = 16.27$$
$$\text{MAPE} = 19.24\%$$

这些预测精度的测度简单地度量了预测方法如何能很好地预测时间序列的历史数值。现在，假设我们想要预测未来时期（如第 13 周）的销售量。在这种情形下，第 13 周的预测值是第 12 周时间序列的实际销售量 22。这是第 13 周销售量的一个准确的估计值吗？很遗憾，没有解决未来时期预测精度的方法。但是，如果我们选择一种适用于历史数据的预测方法，而且我们相信历史模式可以持续到未来，那么我们应该得到最终被证实是好的结果。

㊀ 在回归分析中，残差是指应变量的观测值与估计值之间的差。预测误差与回归分析中的残差相似。
㊁ 在回归分析中，MSE 是残差平方和除以它的自由度。在预测中，MSE 是预测误差平方和的平均数。

在结束本节之前，我们考虑表 17-1 的汽油销售量时间序列的另一种预测方法。假设我们用所有得到的历史数据的平均值作为下一期的预测值。我们开始建立第 2 周的预测值。因为对第 2 周只得到一个历史数据，因此第 2 周的预测值就是第 1 周的时间序列值；即第 2 周的预测值是 17 千加仑。为计算第 3 周的预测值，我们计算第 1 周和第 2 周销售量的平均值。因此

$$第 3 周的预测值 = \frac{17+21}{2} = 19$$

类似地，第 4 周的预测值为

$$第 4 周的预测值 = \frac{17+21+19}{3} = 19$$

用这种方法得到的汽油时间序列的预测值在表 17-8 的预测一列中。利用表 17-8 的结果，我们得到平均绝对误差、均方误差和平均绝对百分数误差的值如下：

$$平均绝对误差 = \frac{26.81}{11} = 2.44$$

$$均方误差 = \frac{89.07}{11} = 8.10$$

$$平均绝对百分数误差 = \frac{141.34}{11} = 12.85\%$$

表 17-8　用所有历史数据平均值作为下一期预测值时，预测及其预测精度测量的计算表

周	时间序列值	预测值	预测误差	预测误差的绝对值	预测误差的平方	百分数误差	百分数误差的绝对值
1	17						
2	21	17.00	4.00	4.00	16.00	19.05	19.05
3	19	19.00	0.00	0.00	0.00	0.00	0.00
4	23	19.00	4.00	4.00	16.00	17.39	17.39
5	18	20.00	-2.00	2.00	4.00	-11.11	11.11
6	16	19.6	-3.60	3.60	12.96	-22.50	22.50
7	20	19.00	1.00	1.00	1.00	5.00	5.00
8	18	19.14	-1.14	1.14	1.31	-6.35	6.35
9	22	19.00	3.00	3.00	9.00	13.64	13.64
10	20	19.33	0.67	0.67	0.44	3.33	3.33
11	15	19.4	-4.4	4.4	18.36	-29.32	29.33
12	22	19.00	3.00	3.00	9.00	13.64	13.64
		合计	4.53	26.81	89.07	2.76	141.34

现在我们通过比较每一种方法的 MAE、MSE 和 MAPE 的值，来比较两种预测方法的精度。

	朴素预测法	过去数值平均数法
MAE	3.73	2.44
MSE	16.27	8.10
MAPE	19.24%	12.85%

对每一种测度，过去数值平均数法都比用最近一期观测值作为下一期预测值的方法提供了更准确的预测。一般地，如果时间序列基本上是平稳的，所有历史数据平均值法将永远提供最好的结果。

假设基本的时间序列是不平稳的，在第 17.1 节我们曾提到，经济条件的改变常常导致有水平模式的时间序列移动到一个新的水平。我们讨论汽油批发商与佛蒙特州警方签署了一份合同，给位于佛蒙特州南部的警车加油。表 17-2 中给出了汽油销售量的原始时间序列以及签署了新合同后 10 周的销售量数据。图 17-2 是相应的时间序列图。注意，由此产生的时间序列第 13 周的水平改变了。当这种移动到新水平的情况出现时，用所有历史数据平均值法调整到时间序列的新水平需要很长时间。在这种情形下，简单的朴素方法对水平的改变调整迅速，因

为它用最近一期的观测值作为预测值。

在比较不同预测方法时，预测精度的测度是一个重要因素。但是，我们需要小心，不能完全依赖它们。当选择预测方法时，对有可能影响预测的经济状况的良好判断和知识也应该仔细考虑。同时历史预测精度不是唯一的考虑因素，时间序列未来有可能改变的情形下尤为如此。

在下一节我们将介绍对呈现水平模式的时间序列进行预测的更完善的方法。用这里介绍的预测精度的测度，我们将确定用下一节介绍方法得到的预测比用本节介绍的简单方法得到的预测更为准确。我们将介绍的这些方法能较好地适应时间序列改变到一个新水平的情形。预测方法快速适应水平改变的能力是重要的考虑因素，尤其是对短期预测情形。

练习

方法

☆2. 参考第 1 题的时间序列数据。用所有历史数据平均值作为下一期预测值。计算下列预测精度的测度。

a. MAE。

b. MSE。

c. MAPE。

d. 第 7 周的预测值是多少？

4. 考虑下列时间序列数据：

月	1	2	3	4	5	6	7
数值	24	13	20	12	19	23	15

a. 计算用最近一期数值作为下一期预测值的 MSE。8 月的预测值是多少？

b. 计算用所有数值的平均值作为下一期预测值的 MSSE。8 月的预测值是多少？

c. 哪种方法提供了较好的预测？

17.3　移动平均法和指数平滑法

本节我们将讨论适合有水平模式的时间序列的 3 种预测方法：移动平均法、加权移动平均法和指数平滑法。这些方法能很好地适应水平模式的水平数值的改变，比如我们看到的汽油销售量的扩展时间序列（表 17-2 和图 17-2）。但是，当有明显的趋势、循环或季节影响时，这些方法未经修正是不适合的。由于每一种方法的目的都是为了“消除”时间序列的随机波动，因此它们都被称为平滑方法。这些方法易于使用，通常对短期预测，例如下一个时期的预测，可以提供较高的精度水平。

17.3.1　移动平均法

移动平均法（moving average）使用时间序列中最近 k 期数据值的平均数作为下一个时期的预测值。在数学上，k 阶移动平均预测如下：

k 阶移动平均预测法

$$F_{t+1} = \frac{\text{最近 } k \text{ 期数据值之和}}{k} = \frac{Y_t + Y_{t-1} + \cdots + Y_{t-k+1}}{k} \tag{17-1}$$

式中，F_{t+1} 代表时间序列 $t+1$ 期的预测值；Y_t 代表时间序列 t 期的实际值。

术语移动的含义是，每次使用时间序列的新观测值，用它代替公式中最旧的观测值，从而计算出一个新的平均数。因此，当使用新观测值时，平均数将会改变或者移动。

为了说明移动平均法，我们回到表 17-1 和图 17-1 的汽油销售量数据。图 17-1 的时间序列图表明汽油销售量时间序列具有水平模式。因此，本节的平滑方法是适用的。

为了用移动平均法来预测时间序列，首先我们应选择移动平均法的阶或所包含的时间序列值的个数。如果仅仅考虑时间序列最近的值被认为是相关的，则应选择较小的 k 值；如果更多过去的值被认为是相关的，则较大的 k 值较好。在前面提到过，具有水平模式的时间序列随时间的推移能移动到一个新的水平。移动平均将适应序列

的新水平，并重新给出 k 个时期的好的预测值。因此，较小的 k 值将更快速追踪时间序列的移动，而较大的 k 值将随着时间的推移更有效地消除随机波动。因此，以对时间序列态势的了解为依据的管理判断，是有助于选择一个合适的 k 值的。

为了说明如何用移动平均法预测汽油的销售量，我们使用3周移动平均（$k=3$）。首先我们用第1~3周的时间序列值的平均数作为第4周销售量的预测值。

$$F_4 = \text{第 1} \sim 3 \text{ 周的平均数} = \frac{17+21+19}{3} = 19$$

因此，第4周汽油销售量的移动平均预测值为19或19 000加仑。因为第4周的实际值为23，所以第4周的预测误差为23－19＝4。

接下来，我们用第2~4周时间序列值的平均数计算第5周销售量的预测值。

$$F_5 = \text{第 2} \sim 4 \text{ 周的平均数} = \frac{21+19+23}{3} = 21$$

因此，第5周销售量的预测值为21，预测误差为17－21＝－3。表17-9给出了汽油销售量时间序列的3周的移动平均预测值的全部汇总，图17-7是原始时间序列图和3周移动平均预测图。注意，移动平均预测图显示出已经消除了时间序列的随机波动。

表17-9　3周移动平均计算的汇总表

周	时间序列值	移动平均预测值	预测误差	预测误差的绝对值	预测误差的平方	百分数误差	百分数误差的绝对值
1	17						
2	21						
3	19						
4	23	19	4	4	16	17.39	17.39
5	18	21	-3	3	9	-16.67	16.67
6	16	20	-4	4	16	-25.00	25.00
7	20	19	1	1	1	5.00	5.00
8	18	18	0	0	0	0.00	0.00
9	22	18	4	4	16	18.18	18.18
10	20	20	0	0	0	0.00	0.00
11	15	20	-5	5	25	-33.33	33.33
12	22	19	3	3	9	13.64	13.64
		合计	0	24	92	-20.79	129.21

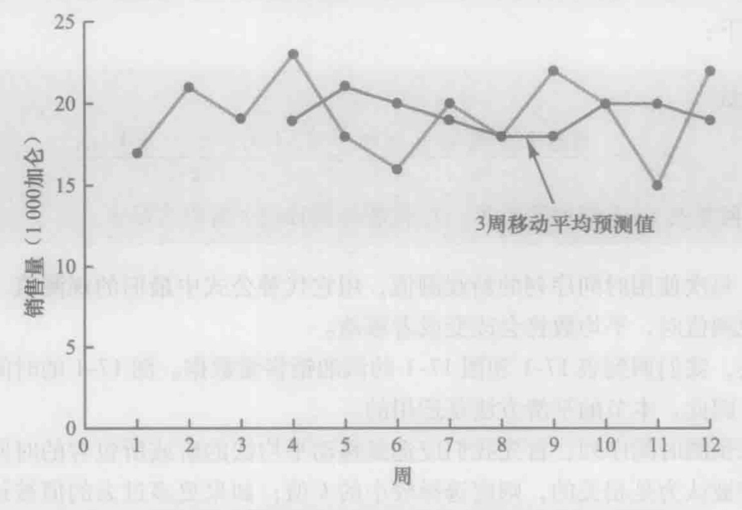


图17-7　汽油销售量的时间序列图和3周的移动平均预测值

为了预测第 13 周，即未来的下一时期的销售量，我们简单地计算第 10 周，第 11 周和第 12 周时间序列的平均数。

$$F_{13} = \text{第 10} \sim \text{12 周的平均数} = \frac{20+15+22}{3} = 19$$

因此，第 13 周的预测值为 19 或 19 000 加仑汽油。

预测精度 在第 17.2 节我们讨论了 3 种预测精度的测度：MAE，MSE 和 MAPE。利用表 17-9 的 3 周移动平均的计算结果，3 种预测精度的测度分别为：

$$\text{MAE} = \frac{24}{9} = 2.67$$

$$\text{MSE} = \frac{92}{9} = 10.22$$

$$\text{MAPE} = \frac{129.21}{9} = 14.36\%$$

在第 17.2 节我们还给出了用最近一期的观测值作为下一周的预测值（$k=1$ 阶的移动平均）的结果为：MAE = 3.73，MSE = 16.27 和 MAPE = 19.24% ㊀。因此，在每一种情况下，3 周移动平均方法都比简单地用最近的观测值作为预测值给出了更准确的预测。

为了确定使用不同的阶数 k 的移动平均数是否可以给出更准确的预测，我们建议用试验和误差来确定使 MSE 达到最小的 k 值。对汽油销售量时间序列，能够证明，MSE 的最小值对应于 $k=6$ 阶的移动平均，其 MSE = 6.79。如果我们愿意假定，对历史数据最合适的移动平均数的阶数对未来值也是最合适的，则用 $k=6$ 阶的移动平均可以得到汽油销售量最准确地移动平均预测。

17.3.2 加权移动平均法

在移动平均法中，移动平均数计算中的每个观测值都使用相同的权重。一种称为**加权移动平均**（weighted moving average）的方法对此做了改变，即对每个数据值选择不同的权重，然后计算最近 k 期数据值的加权平均数作为预测值。在大多数情况下，最近时期的观测值得到最大的权重，而减少较远期的观测值的权重。让我们用汽油销售量时间序列来说明 3 周加权移动平均数的计算。我们指定最近时期观测值的权重为 3/6，第二近的观测值的权重为 2/6，第三近的观测值的权重为 1/6。用加权移动平均，第 4 周的预测值计算如下：

$$\text{第 4 周的预测值} = \frac{1}{6} \times 17 + \frac{2}{6} \times 21 + \frac{3}{6} \times 19 = 19.33$$

注意，加权移动平均方法的权重之和应该等于 1 ㊁。

预测精度 为了使用加权移动平均法，我们首先应该选择加权移动平均数中所包含的数据值个数，然后对每个数据值选择权重。在一般情形下，如果我们相信，最近的数据比过去的数据能更好地预测未来，则我们应该给较近的观测值更大的权重。但是，当时间序列的波动较大时，对每个数据值选择近似相等的权重也许是最好的方法。选择权重的唯一要求是它们的总和必须等于 1。为了确定数据值个数和权重的一个特殊组合是否比其他组合提供了更精确的预测，我们建议使用 MSE 作为预测精度的测度。即如果我们假定能最好地反映过去的这个组合也能最好地反映未来，我们将使用对历史时间序列的 MSE 达到最小的数据值个数和权重的组合，来预测时间序列下一个时期的数值。

17.3.3 指数平滑法

指数平滑法（exponential smoothing）㊂也是利用过去的时间序列值的加权平均数作为预测值；它是加权移动平

㊀ 在需要比较不同时间长度的预测方法时，如比较周销售量的预测与月销售量的预测，相对测度 MAPE 是首选。

㊁ $k=3$ 阶的移动平均预测恰好是加权移动平均的特例，其每个权重都是 1/3。

㊂ 还有一些指数平滑法。这里介绍的方法常常被称为单一指数平滑法。在下节我们将演示如何用两个平滑常数的指数平滑法来预测有线性趋势的时间序列。

均法的一个特例，即我们只选择一个权重——最近时期观测值的权重。其他数据值的权重可以自动推算出来，并且随着观测值距离预测期越远，权重也变得越小。指数平滑法的基本方程如下：

指数平滑预测

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t \tag{17-2}$$

式中，F_{t+1}代表时间序列 $t+1$ 期的预测值；Y_t代表时间序列 t 期的实际值；F_t代表时间序列 t 期的预测值；α 代表平滑常数（$0 \leqslant \alpha \leqslant 1$）。

式（17-2）表明 $t+1$ 期的预测值是 t 期的实际值和预测值的加权平均数，t 期实际值的权重为**平滑常数**（smoothing constant）α，t 期预测值的权重为 $1-\alpha$。可以证明，任何时期的指数平滑预测值实际上是时间序列所有过去实际数据的一个加权平均数。让我们通过一个仅包含三个时期数据 Y_1，Y_2 和 Y_3 的时间序列来说明。

开始计算，我们令 F_1 等于时间序列 1 期的实际值，即 $F_1 = Y_1$。因此，2 期的预测值为

$$F_2 = \alpha Y_1 + (1-\alpha)F_1 = \alpha Y_1 + (1-\alpha)Y_1 = Y_1$$

我们看到，2 期的指数平滑预测值等于时间序列 1 期的实际值。

3 期的预测值为

$$F_3 = \alpha Y_2 + (1-\alpha)F_2 = \alpha Y_2 + (1-\alpha)Y_1$$

最后，将 F_3 的表达式代入 F_4 的表达式中，我们得到

$$\begin{aligned} F_4 &= \alpha Y_3 + (1-\alpha)F_3 = \alpha Y_3 + (1-\alpha)[\alpha Y_2 + (1-\alpha)Y_1] \\ &= \alpha Y_3 + \alpha(1-\alpha)Y_2 + (1-\alpha)^2 Y_1 \end{aligned}$$

现在我们看到，F_4 是前三个时间序列数值的加权平均数，Y_1，Y_2 和 Y_3 的系数或权重之和等于 1。在一般情形下，一个类似的结论可以表述为：任何预测值 F_{t+1}是所有时间序列过去数值的加权平均数[⊖]。

尽管指数平滑法提供的预测值是所有过去观测值的加权平均数，但是并不需要保存过去所有的数据，以计算下一个时期的预测值。事实上，式（17-2）表明，一旦选定了平滑常数 α 的值，我们仅仅需要两项的信息就可计算预测值：t 期时间序列的实际值 Y_t 和 t 期的预测值 F_t。

为了说明指数平滑法，让我们再次考虑表 17-1 和图 17-1 中的汽油销售量时间序列。如上所述，开始计算时，我们设 2 期的指数平滑预测值等于时间序列 1 期的实际值。于是，因为 $Y_1 = 17$，我们得到 $F_2 = 17$。参考表 17-1 的时间序列数据，我们发现时间序列 2 期的实际值 $Y_2 = 21$。因此，2 期的预测误差为 $21 - 17 = 4$。

利用平滑常数 $\alpha = 0.2$ 继续进行指数平滑计算。我们可以得到 3 期的预测值如下：

$$F_3 = 0.2 \times Y_2 + 0.8 \times F_2 = 0.2 \times 21 + 0.8 \times 17 = 17.8$$

一旦得到时间序列 3 期的实际值 $Y_3 = 19$，我们就可以得到 4 期的预测值如下：

$$F_4 = 0.2 \times Y_3 + 0.8 \times F_3 = 0.2 \times 19 + 0.8 \times 17.8 = 18.04$$

继续进行指数平滑计算，我们得到表 17-10 所示的周预测值。注意，我们没有给出第 1 周指数平滑预测值或预测误差，这是因为没有办法做出预测。对第 12 周，我们有 $Y_{12} = 22$，$F_{12} = 18.48$。我们能利用这些信息得到第 13 周的预测值为

$$F_{13} = 0.2 \times Y_{12} + 0.8 \times F_{12} = 0.2 \times 22 + 0.8 \times 18.48 = 19.18$$

因此，第 13 周销售量的指数平滑预测值为 19.18 千加仑或 19 180 加仑的汽油。根据这个预测值，公司可以制定计划和作出决定。

图 17-8 是时间序列的实际值和预测值的图。尤其要注意的是，预测值“消除”了时间序列中的不规则或

⊖ 指数平滑术语来自对历史数据进行加权的指数性质。

随机波动。

表 17-10　平滑常数 $\alpha=0.2$ 的汽油销售量时间序列的指数平滑预测和预测误差的汇总表

周	时间序列值	预测值	预测误差	预测误差的平方
1	17			
2	21	17.00	4.00	16.00
3	19	17.80	1.20	1.44
4	23	18.04	4.96	24.60
5	18	19.03	−1.03	1.06
6	16	18.83	−2.83	8.01
7	20	18.26	1.74	3.03
8	18	18.61	−0.61	0.37
9	22	18.49	3.51	12.32
10	20	19.19	0.81	0.66
11	15	19.35	−4.35	18.92
12	22	18.48	3.52	12.39
		合计	10.92	98.80

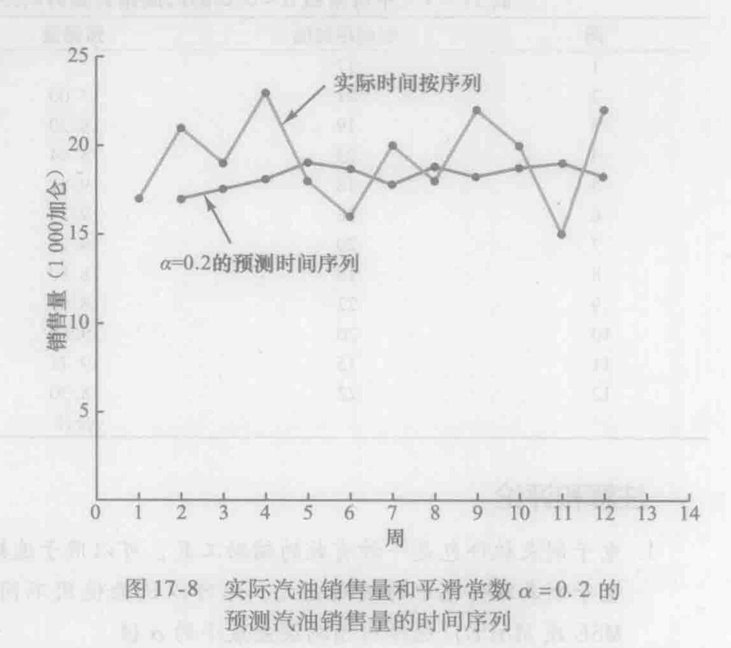


图 17-8　实际汽油销售量和平滑常数 $\alpha=0.2$ 的预测汽油销售量的时间序列

预测精度　在前面的指数平滑计算中，我们使用的平滑常数 $\alpha=0.2$。尽管 α 取 0 和 1 之间的任何值都是可以接受的，但是有些 α 值得到的预测值比其他 α 值更好。为了观察如何得到一个合适的 α 值，将基本指数平滑模型改写如下：

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t$$

$$F_{t+1} = \alpha Y_t + F_t - \alpha F_t$$

$$F_{t+1} = F_t + \alpha(Y_t - F_t) \tag{17-3}$$

因此，新的预测值 F_{t+1} 等于过去的预测值 F_t 加上一个调整值，这个调整值等于 α 乘以最近时期的预测误差 $(Y_t - F_t)$。即通过调整 t 期的预测值和一部分预测误差，我们就可以得到 $t+1$ 期的预测值。如果时间序列包含大量的随机波动，我们倾向于选择较小的平滑常数。这样选择的原因是：大多数预测误差是由随机波动引起的，我们不希望对预测做出过度的反应和过快的调整。对随机波动相对较小的时间序列，预测误差更可能表示序列水平的改变。于是，选用较大的平滑常数的优点是迅速调整预测，这使得预测对改变的条件反应更迅速。

我们用来确定平滑常数 α 合理值的准则，与我们建议的确定所包含在移动平均数计算中的阶数或数据时期个数的准则相同。也就是说，我们选择 MSE 达到最小的 α 值。对于 $\alpha=0.2$ 的汽油销售量指数平滑预测，表 17-10 给出了 MSE 的计算结果汇总。注意，预测误差平方项比周数少一项，这是因为我们没有过去的数据作为 1 期的预测值。预测误差平方和的值为 98.80，因此 MSE = 98.80/11 = 8.98。就较小的 MSE 而言，不同的 α 值能否提供一个较好的结果呢？也许回答这个问题最简单的方法是用另一个 α 值来试验，然后将所得到的 MSE 与用平滑常数 $\alpha=0.2$ 得到的 MSE = 8.98 进行比较。

表 17-11 是用 $\alpha=0.3$ 得到的指数平滑预测结果，预测误差平方和的值为 102.83，因此 MSE = 102.83/11 = 9.35。由于 MSE = 9.35，我们看到，对当前数据集，使用平滑常数 $\alpha=0.3$ 得到的结果比使用平滑常数 $\alpha=0.2$ 得到的结果的预测精度要差。于是，我们倾向于选择原来的平滑常数 $\alpha=0.2$。对其他的 α 值，利用反复试验的计算方法，我们可以找到一个“好”的平滑常数值。这个数值可以用于指数平滑模型来预测未来。以后的数据，即我们得到时间序列的新的观测值以后，将分析这个新搜集的时间序列数据，决定是否应调整平滑常数以得到更好的预测结果。

表 17-11 平滑常数 $\alpha=0.3$ 的汽油销售量时间序列的指数平滑预测和预测误差的汇总表

周	时间序列值	预测值	预测误差	预测误差的平方
1	17			
2	21	17.00	4.00	16.00
3	19	18.20	0.80	0.64
4	23	18.44	4.56	20.79
5	18	19.81	−1.81	3.28
6	16	19.27	−3.27	10.69
7	20	18.29	1.71	2.92
8	18	18.80	−0.80	0.64
9	22	18.56	3.44	11.83
10	20	19.59	0.41	0.17
11	15	19.71	−4.71	22.18
12	22	18.30	3.70	13.69
		合计	8.03	102.83

注释和评论

1. 电子制表软件包是一种有效的辅助工具，可以用于选择指数平滑法中合适的 α 值。利用时间序列的数据和电子制表软件包中的预测公式，你可以试验使用不同的 α 值，并用一个或多个预测精度的测度（MAE、MSE 或 MAPE）选择出预测误差最小的 α 值。
2. 我们在平稳时间序列的背景下讨论移动平均法和指数平滑法。这些方法也可以用于水平改变但没有呈现趋势或季节的非平稳时间序列的预测。较小 k 值的移动平均比较大 k 值的移动平均适应得更快。平滑常数接近于 1 的指数平滑模型比较小平滑常数的模型适应得更快。

练习

方法

6. 考虑下列时间序列数据：

月	1	2	3	4	5	6	7
数值	24	13	20	12	19	23	15

绘制时间序列图，数据中呈现何种类型的模式？

a. 计算时间序列 3 个月移动平均预测值，计算 MSE 和 8 月的预测值。

b. 计算时间序列 $\alpha=0.2$ 的指数平滑预测值。计算 MSE 和 8 月的预测值。

c. 比较 3 个月移动平均法和 $\alpha=0.2$ 的指数平滑法，基于 MSE，哪种方法给出了更准确的预测？

d. 计算平滑常数 $\alpha=0.4$ 的指数平滑预测值。基于 MSE，平滑常数 0.2 还是 0.4 给出了更准确的预测？请解释。

8. 再次参见表 17-1 中的汽油销售量的时间序列。

a. 令 1/2，1/3 和 1/6 分别为最近的、第二近和第三近的观测值的权重，计算时间序列的 3 周加权移动平均数。

b. 计算（a）的加权移动平均数的 MSE。在加权移动平均数和不加权移动平均数中，你更倾向于哪一种方法？不加权移动平均的 MSE 为 10.22。

c. 假设允许你选择任何总和为 1 的权重。你是否总能找到这样一组权重，使得加权移动平均法的 MSE 指数比不加权移动平均法的 MSE 要小？请解释。

10. 对表 17-1 的数据，使用平滑常数 $\alpha=0.2$，式（17-2）表明第 13 周汽油销售量的预测值为 $F_{13}=0.2Y_{12}+0.8F_{12}$。然而，第 12 周的预测值为 $F_{12}=0.2Y_{11}+0.8F_{11}$。因此，将这两个结果合并起来，第 13 周的预测值可以写为

$$F_{13}=0.2Y_{12}+0.8\times(0.2Y_{11}+0.8F_{11})$$
$$=0.2Y_{12}+0.16Y_{11}+0.64F_{11}$$

a. 根据 $F_{11}=0.2Y_{10}+0.8F_{10}$（对 F_{10} 和 F_{9} 类似），继续扩展 F_{13} 的表达式，直到它表示为过去数据值 Y_{12}，Y_{11}，Y_{10}，Y_{9}，Y_{8} 和第 8 期预测值为止。

b. 参阅过去数据值 Y_{12}，Y_{11}，Y_{10}，Y_{9} 和 Y_{8} 的系数或权重。对指数平滑法在得到新预测值时，是如何对过去数值分配权重的，你有什么样的观测结果？请将这个加权形式与移动平均法的加

权形式进行比较。

应用

12. 下面是某个公司AAA级债券在连续12个月中的利率。

9.5 9.3 9.4 9.6 9.8 9.7 9.8

10.5 9.9 9.7 9.6 9.6

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 计算时间序列的3个月和4个月的移动平均数。基于MSE，3个月或4个月的移动平均数是否提供准确的预测？请解释。

c. 下一个月的移动平均预测值是多少？

14. 下面是某种特定产品在过去12个月中销售量的时间序列。

月	销售量	月	销售量
1	105	7	145
2	135	8	140
3	120	9	100
4	105	10	80
5	90	11	100
6	120	12	110

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 计算时间序列的$\alpha=0.3$的指数平滑预测值。

c. 计算平滑常数$\alpha=0.5$的指数平滑预测值。基于MSE，平滑常数0.3和0.5哪个提供更准确的预测？

16. 美国人口普查局追踪每年各月新房销售价格的中位数。下面是1990～2011年4月价格的中位数（U. S. Census Bureau website，2012年4月16日）。

年份	价格（1 000美元）	年份	价格（1 000美元）
1990	130.0	2001	175.2
1991	121.0	2002	187.1
1992	120.0	2003	189.5
1993	127.0	2004	222.3
1994	129.0	2005	236.3
1995	134.0	2006	257.0
1996	140.0	2007	242.5
1997	150.0	2008	246.4
1998	148.0	2009	219.2
1999	159.9	2010	208.3
2000	162.6	2011	224.7

a. 绘制时间序列图，数据中呈现何种类型的模式？针对这个时间序列，对有可能导致时间序列图模式的一些因素进行讨论。

b. 指出（a）中绘制的时间序列图的模式，你是否认为本节所讨论的预测方法是计算预测值的适合方法？请解释。

c. 你会使用过去多少数据来预测2012年4月的数值？请解释。

17.4 趋势推测法

本节我们介绍三种适合于具有趋势模式的时间序列的预测方法。首先，我们说明如何用简单线性回归预测具有线性趋势的时间序列。然后，我们解释如何用Holt线性指数平滑法进行预测，它是单一指数平滑法的一个推广，使用两个平滑常数：一个用于解释时间序列的水平，另一个用于解释数据中的线性趋势。最后，我们演示如何用回归分析的曲线拟合功能预测具有曲线或非线性趋势的时间序列。

17.4.1 线性趋势回归

在第17.1节中我们用表17-3和图17-3中的自行车销售量的时间序列，来说明具有趋势模式的时间序列。让我们用这个时间序列来解释如何用简单线性回归预测具有线性趋势的时间序列。自行车销售量的时间序列的数据再一次出现在表17-12和图17-9中。

表17-12 自行车销售量时间序列

年	销售量（1 000辆）	年	销售量（1 000辆）	年	销售量（1 000辆）
1	21.6	5	23.9	9	28.6
2	22.9	6	27.5	10	31.4
3	25.5	7	31.5		
4	21.9	8	29.7		

尽管图17-9的时间序列图显示在过去10年有一些上下波动，但我们可能会同意，图17-10中所示的线性趋势对时间序列的长期变动提供了一个合理的近似。我们用简单线性回归的方法（见第14章）为自行车销售量时

间序列建立一条线性趋势线。

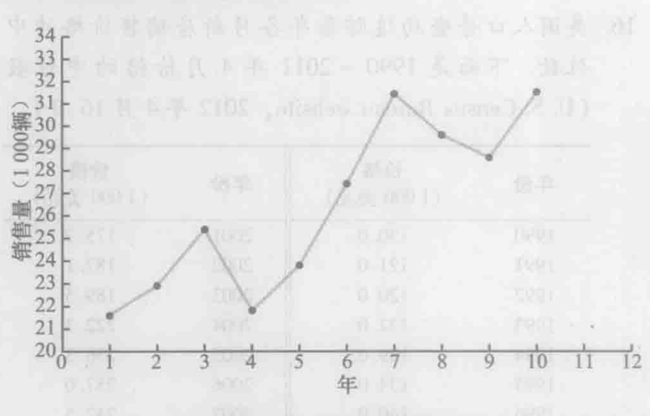


图 17-9 自行车销售量时间序列

图 17-10 用线性函数描述自行车销售量的趋势

在第 14 章，描述自变量 x 和应变量 y 之间直线关系的估计回归方程为

$$\hat{y} = b_0 + b_1 x$$

式中，$\hat{y}$ 是 y 的估计值或预测值，为了强调在这个预测中自变量是时间这一事实，我们将用 t 来代替 x，用 T_t 来代替 $\hat{y}$，以强调我们估计的是时间序列的趋势。因此，对时间序列的线性趋势，我们用如下的估计回归方程：

线性趋势方程

$$T_t = b_0 + b_1 t \tag{17-4}$$

式中，T_t 为 t 期线性趋势预测值；b_0 为线性趋势线的截距；b_1 为线性趋势线的斜率；t 为期间。

在式（17-4）中，时间变量初始值 $t=1$ 对应于时间序列的第一个观察值（自行车销售量时间序列的第 1 年），$t=n$ 对应于时间序列的最近观察值（自行车销售量时间序列的第 10 年）。因此，对自行车销售量时间序列，$t=1$ 对应于最早的时间序列值，而 $t=10$ 对应于最近的年份。

式（17-4）中的估计回归系数（b_0 和 b_1）的计算公式如下：

计算线性趋势线的斜率和截距的公式[⊖]

$$b_1 = \frac{\sum_{t=1}^{n}(t-\bar{t})(Y_t-\bar{Y})}{\sum_{t=1}^{n}(t-\bar{t})^2} \tag{17-5}$$

$$b_0 = \bar{Y} - b_1\bar{t} \tag{17-6}$$

式中，Y_t 为 t 期时间序列值；n 代表时期个数（观测值的个数）；$\bar{Y}$为时间序列的平均值；$\bar{t}$为 t 的平均值。

为了计算自行车销售量时间序列的线性趋势方程，我们利用表 17-12 中的信息，计算$\bar{t}$和$\bar{Y}$。

$$\bar{t} = \frac{\sum_{t=1}^{n} t}{n} = \frac{55}{10} = 5.5 \qquad \bar{Y} = \frac{\sum_{t=1}^{n} Y_t}{n} = \frac{264.5}{10} = 26.45$$

利用这些数值和表 17-13 中的信息，我们可以计算自行车销售量时间序列的斜率和截距。

⊖ b_1 的另一个公式是

$$b_1 = \frac{\sum_{t=1}^{n} tY_t - \left(\sum_{t=1}^{n} t \sum_{t=1}^{n} Y_t\right)/n}{\sum_{t=1}^{n} t^2 - \left(\sum_{t=1}^{n} t\right)^2/n}$$

当使用计算器计算 b_1 时，往往会推荐式（17-5）的形式。

表 17-13 自行车销售量时间序列的线性趋势计算汇总表

	t	Y_t	$t-\bar{t}$	$Y_t-\bar{Y}$	$(t-\bar{t})(Y_t-\bar{Y})$	$(t-\bar{t})^2$
	1	21.6	-4.5	-4.85	21.825	20.25
	2	22.9	-3.5	-3.55	14.425	12.25
	3	25.5	-2.5	-0.95	2.375	6.25
	4	21.9	-1.5	-4.55	6.825	2.25
	5	23.9	-0.5	-2.55	1.275	0.25
	6	27.5	0.5	1.05	0.525	0.25
	7	31.5	1.5	5.05	7.575	2.25
	8	29.7	2.5	3.25	8.125	6.25
	9	28.6	3.5	2.15	7.525	12.25
	10	31.4	4.5	4.95	22.275	20.25
合计	55	264.5			90.750	82.50

$$b_1 = \frac{\sum_{t=1}^{n}(t-\bar{t})(Y_t-\bar{Y})}{\sum_{t=1}^{n}(t-\bar{t})^2} = \frac{90.75}{82.5} = 1.1$$

$$b_0 = \bar{Y} - b_1\bar{t} = 26.45 - 1.1 \times 5.5 = 20.4$$

因此，线性趋势方程为

$$T_t = 20.4 + 1.1t$$

斜率 1.1 表明在过去的 10 年中，工厂的销售量每年平均增长大约为 1 100 个单位。如果我们假定过去 10 年的销售量趋势对未来也是合适的，则趋势方程可用来预测时间序列的趋势成分。例如，将 $t=11$ 代入方程便可以得到下一年的趋势预测值 T_{11}，即

$$T_{11} = 20.4 + 1.1 \times 11 = 32.5$$

因此，利用趋势推测，我们将预测出下一年的自行车销售量为 32 500 辆。

为了计算趋势推测预测方法的精度，我们将使用均方误差 MSE。表 17-14 给出了自行车销售量时间序列误差平方和的计算结果。因此，自行车销售量时间序列，

$$\text{MSE} = \frac{\sum_{t=1}^{n}(Y_t - F_t)^2}{n} = \frac{30.7}{10} = 3.07$$

表 17-14 自行车销售量时间序列的线性趋势预测和预测误差的汇总表

年	销售量（1 000 辆）Y_t	预测值 F_t	预测误差	预测误差的平方
1	21.6	21.5	0.1	0.01
2	22.9	22.6	0.3	0.09
3	25.5	23.7	1.8	3.24
4	21.9	24.8	-2.9	8.41
5	23.9	25.9	-2.0	4.00
6	27.5	27.0	0.5	0.25
7	31.5	28.1	3.4	11.56
8	29.7	29.2	0.5	0.25
9	28.6	30.3	-1.7	2.89
10	31.4	31.4	0.0	0.00
			合计	30.70

在预测时，由于线性趋势回归使用与第 14 章介绍的相同回归分析程序，我们可以利用 Minitab 或 Excel 的常规回归分析程序完成计算。图 17-11 是用 Minitab 回归分析模块得到的自行车销售量时间序列的计算机输出结果。

```
The regression equation is
Y = 20.4 + 1.10 t

Predictor        Coef        SE Coef        T        p
Constant         20.400      1.338          15.24    0.000
t                1.100 0     0.215 7        5.10     0.001

S = 1.958 95          R-sq = 76.5%          R-sq (adj) = 73.5%

Analysis of Variance

SOURCE            DF     SS         MS        F        p
Regression        1      99.825     99.825    26.01    0.001
Residual Error    8      30.700     3.837
Total             9      130.525
```

图 17-11　自行车销售量时间序列的 Minitab 回归输出结果

在图 17-11 的方差分析（ANOVA）表中，MSE 的值为

$$\text{MSE}=\frac{\text{误差平方和}}{\text{自由的}}=\frac{30.7}{8}=3.837$$

这个 MSE 值与我们前面计算的 MSE 值不同，这是因为误差平方和除以 8 而不是 10；于是，在回归输出中 MSE 不是预测误差平方的平均数。然而，大多数预测软件包，计算 MSE 采用误差平方的平均数。因此，当我们用时间序列软件包建立趋势方程时，报告的 MSE 值可能与你利用一般回归方法得到的值略有不同。例如，在图 17-12中，我们显示了用 Minitab's Trend Analysis 时间序列方法得到计算机输出的图示部分，注意 MSD = 3.07 是预测误差平方的平均数[⊖]。

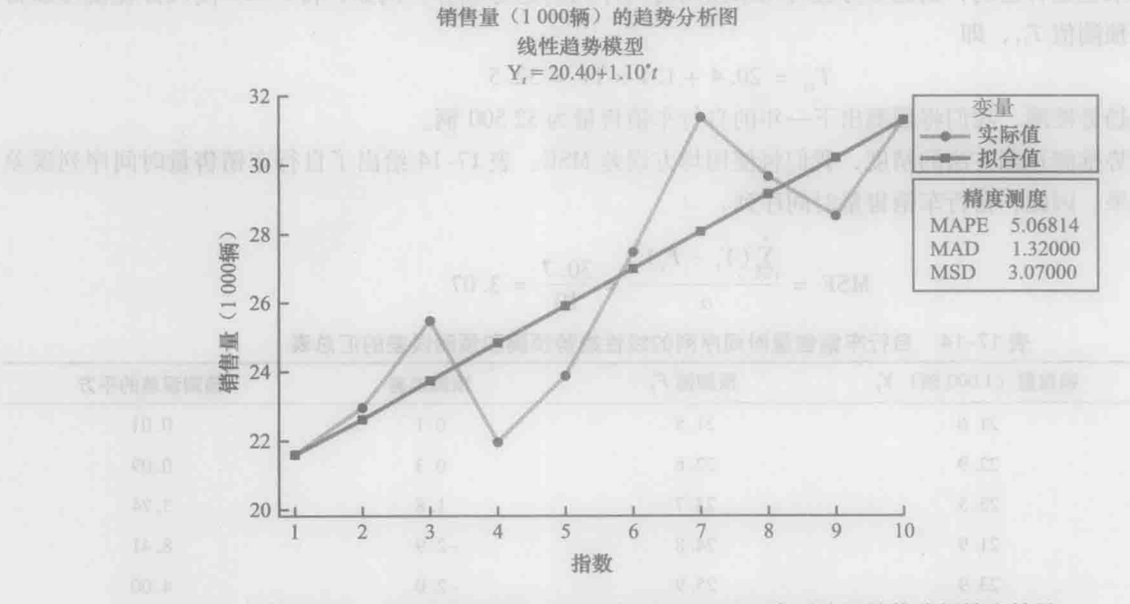


图 17-12　自行车销售量时间序列的 Minitab 时间序列线性趋势分析输出结果

17.4.2 Holt 线性指数平滑

查尔斯·霍尔特（Charles Holt）建立了用于预测具有线性趋势的时间序列的指数平滑形式。回顾第 17.3 节

⊖ 在 Minitab's Trend Analysis 输出中，MSD 是离差平方的平均数，即预测误差平方的平均数。

讨论的指数平滑方法，它使用平滑常数 α“消除”时间序列中随机的或不规则的影响，并且用方程

$$F_{t+1} = \alpha Y_t + (1-\alpha) F_t$$

得到 $t+1$ 期的预测值。

Holt **线性指数平滑**[⊖]（linear exponential smoothing）方法用两个平滑常数 α 和 β 和三个方程得到预测。

Holt 线性指数平滑方程

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + b_{t-1}) \tag{17-7}$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)(b_{t-1}) \tag{17-8}$$

$$F_{t+k} = L_t + b_t k \tag{17-9}$$

式中，L_t 是 t 期时间序列水平的估计值；b_t 是 t 期时间序列斜率的估计值；α 是时间序列水平的平滑常数；β 是时间序列斜率的平滑常数；F_{t+k} 是向前 k 期的预测值；k 是向前预测的时期个数。

让我们对表 17-12 中的自行车销售量时间序列，应用 $\alpha=0.1$，$\beta=0.2$ 的 Holt 方法。为了启动该方法，我们需要时间序列第 1 年的水平估计值 L_1 和时间序列第 1 年的斜率估计值 b_1。一般方法是令 $L_1=Y_1$，$b_1=Y_2-Y_1$。利用此启动程序，我们得到

$$L_1 = Y_1 = 21.6$$

$$b_1 = Y_2 - Y_1 = 22.9 - 21.6 = 1.3$$

对 $k=1$，利用式（17-9）计算第 2 年销售量的预测值为 $F_2 = L_1 + b_1 = 21.6 + 1.3 \times 1 = 22.9$。然后，我们继续利用式（17-7）~式（17-9），计算第 2 年水平和趋势的估计值以及第 3 年的预测值。

首先，利用式（17-7）和平滑常数 $\alpha=0.1$，计算时间序列第 2 年的水平估计值。

$$L_2 = 0.1 \times 22.9 + 0.9 \times (21.6 + 1.3) = 22.9$$

注意 $21.6+1.3$ 是第 2 年销售量的预测值。因此，利用式（17-7）得到时间序列第 2 年的水平估计值，是第 2 年的观测值（用权重 $\alpha=0.1$）和第 2 年的预测值（用权重 $1-\alpha=1-0.1=0.9$）的加权平均数。一般地，对观测值（Y_t）的权重更多地给以较大的 α 值，而对预测值（$L_{t-1}+b_{t-1}$）的权重更多地给予较小的值。

接下来我们利用式（17-8）和平滑常数 $\beta=0.2$，计算时间序列第 2 年的斜率估计值

$$b_2 = 0.2 \times (22.9 - 21.6) + (1 - 0.2) \times 1.3 = 1.3$$

时间序列第 2 年的斜率估计值是时间序列第 2 年与第 1 年水平的估计值之差（用权重 $\beta=0.2$）和第 1 年的斜率估计值（用权重 $1-\beta=1-0.2=0.8$）的加权平均数。一般地，对估计的水平之差的权重更多地给予较大的 β 值，而对过去时期斜率估计值的权重更多地给予较小的值。

利用刚得到 L_2、b_2 的估计值及式（17-9），第 3 年销售量的预测值为

$$F_3 = L_2 + b_2 = 22.9 + 1.3 \times 1 = 24.2$$

用类似的方式进行计算，结果显示在表 17-15 中。预测误差的平方和为 39.678，因此 MSE = 39.678/9 = 4.41。

平滑常数 α 和 β 不同的值能提供更准确的预测吗？为了回答这个问题，我们将尝试 α 和 β 的不同组合来确定，是否可以找到使 MSE 的值小于 4.41（用平滑常数 $\alpha=0.1$ 和 $\beta=0.2$ 得到的 MSE 值）的一个组合。可以通过反复测试或者使用有最佳平滑常数组合选项的更高级的统计软件包来寻找 α 和 β 合适的值。

注意第 10 年时间序列的水平估计值是 $L_{10}=32.220$，第 10 年的斜率估计值为 $b_{10}=1.171$。如果我们假定过去 10 年销售量的趋势对未来是合适的，则可以用式（18-9）建立未来时期的预测值。例如，将 $t=11$ 代入式（17-9），得到下一年的趋势推测值或预测值 F_{11}。

⊖ Holt 线性指数平滑常常被称为二重指数平滑。

表 17-15 自行车销售量时间序列的 Holt 线性指数平滑法的计算汇总表
($\alpha=0.1$ 和 $\beta=0.2$)

年	销售量（1 000 辆）Y_t	水平估计值 L_t	斜率估计值 b_t	预测值 F_t	预测误差	预测误差的平方
1	21.6	21.600	1.300			
2	22.9	22.900	1.300	22.900	0.000	0.000
3	25.5	24.330	1.326	24.200	1.300	1.690
4	21.9	25.280	1.251	25.656	−3.756	14.108
5	23.9	26.268	1.198	26.531	−2.631	6.924
6	27.5	27.470	1.199	27.466	0.034	0.001
7	31.5	28.952	1.256	28.669	2.831	8.016
8	29.7	30.157	1.245	30.207	−0.507	0.257
9	28.6	31.122	1.189	31.402	−2.802	7.851
10	31.4	32.220	1.171	32.311	−0.911	0.830
					合计	39.678

$$F_{11} = L_{10} + b_{10} \times 1 = 32.220 + 1.171 = 33.391$$

于是，用 Holt 线性指数平滑法我们得到下一年自行车销售量的预测值为 33 391 辆。

17.4.3 非线性趋势回归

用线性函数来拟合趋势是很常见的。然而，正如我们前面讨论的，有时时间序列有曲线或非线性趋势。作为一个例子我们考虑胆固醇药物前 10 年的销售收入（单位：100 万美元）。表 17-16 是时间序列，图 17-13 是相应的时间序列图。例如，第 1 年的收入是 23.1 百万美元；第 2 年的收入是 21.3 百万美元，依此类推。时间序列图显示有一个总体增长或向上的趋势，但是与自行车销售量时间序列不同，线性趋势似乎并不合适。相反，在拟合长期趋势时，似乎需要一个曲线函数。

表 17-16 胆固醇药物销售收入的时间序列

年（t）	销售收入（100 万美元）
1	23.1
2	21.3
3	27.4
4	34.6
5	33.8
6	43.2
7	59.5
8	64.4
9	74.2
10	99.3

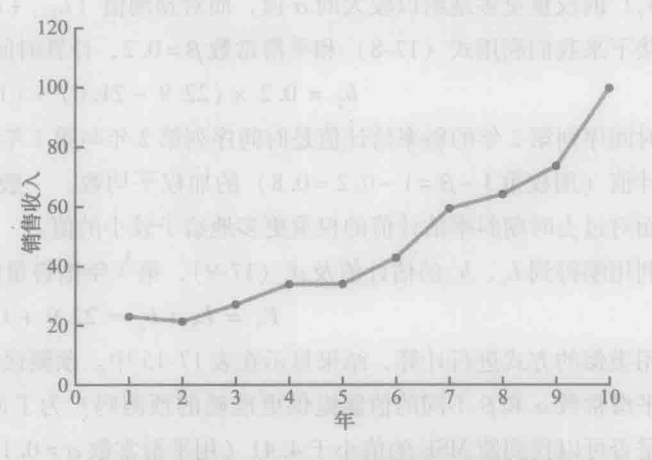


图 17-13 胆固醇药物销售收入的时间序列图（100 万美元）

二次趋势方程（quadratic trend equation） 各种非线性函数能用来建立胆固醇时间序列的趋势估计。例如，考虑下面的二次趋势方程：

$$T_t = b_0 + b_1 t + b_2 t^2 \tag{17-10}$$

对于胆固醇时间序列，$t=1$ 对应于第 1 年，$t=2$ 对应于第 1 年，依此类推。

在第 16.1 节讨论的一般线性模型可以用于计算 b_0，b_1 和 b_2 的值。有两个自变量：年份和年份的平方，应变

量是以 100 万美元计的销售收入。因此，第一个观测值是（1，1，23.1）；第二个观测值是（2，4，21.3），第三个观测值是（3，9，27.4），以此类推。图 17-14 显示了二次趋势模型的 Minitab 多元回归输出结果；估计的回归方程为

$$\text{销售收入(100 万美元)} = 24.2 - 2.11\text{Year} + 0.922\text{YearSq}$$

式中，Year 为 1，2，3，…，10；YearSq 为 1，4，9，…，100。

```
The regression equation is
Revenue = 24.2-2.11 Year + 0.922 YearSq

Predictor       Coef      SE Coef       T         p
Constant      24.182       4.676      5.17      0.001
Year          -2.106       1.953     -1.08      0.317
YearSq         0.921 6     0.173 0    5.33      0.001

S=3.97578 R-Sq=98.1%    R-Sq(adj)=97.6%

Analysis of Variance

SOURCE            DF        SS       MS         F        p
Regression        2      5 770.1   288 5.1   182.52   0.000
Residual Error    7        110.6     15.8
Total             9      5 880.8
```

图 17-14 胆固醇药物销售收入时间序列的 Minitab 二次趋势回归的输出结果

使用标准多元回归程序需要我们计算年份的平方作为第二个自变量值。或者，我们利用 Minitab 的 Time Series—Trend Analysis 程序得到相同的结果。这样做不需要计算年份的平方而且易于使用。当解决涉及二次趋势的练习题时，我们推荐使用这种方法。

指数趋势方程（exponential trend equation） 另一种用于模拟胆固醇时间序列的非线性模式，是用指数模型来拟合这些数据。例如，考虑如下指数趋势方程：

$$T_t = b_0(b_1)^t \tag{17-11}$$

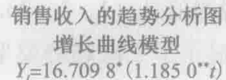
为了更好地理解这个指数趋势方程，假设 $b_0=20$ 和 $b_1=1.2$，然后，对于 $t=1$，$T_1=20\times1.2^1=24$；对于 $t=2$，$T_1=20\times1.2^2=28.8$；对于 $t=3$，$T_1=20\times1.2^3=34.56$。注意，T_t 不像在线性趋势模型中那样以一个固定总量增长，而是以一个固定的百分比增长，增长的百分比为 20%。

Minitab 在它的时间序列模块中有计算指数趋势方程并用这个方程进行预测的功能。不幸的是，Excel 没有这个功能。但在第 16.1 节我们描述了如何对式（17-11）进行对数变换，一般的线性模型方法可以用来计算指数趋势方程。

Minitab 的时间序列模块可以非常容易地用来建立指数趋势方程，不需要涉及对数变换和利用回归分析计算指数趋势方程。图 17-15，我们显示了用 Minitab 的 Time Series—Trend Analysis 程序来拟合指数趋势方程而得到计算机输出的图示部分。

注释和评论

线性趋势回归的基础是找到使预测误差平方和，从而使 MSE 达到最小的估计的回归方程，所以就 MSE 而言，我们期待线性趋势回归的表现优于 Holt 线性指数平滑。例如，自行车销售量时间序列，用线性趋势回归的 MSE = 3.07，而用 Holt 线性指数平滑的 MSE = 3.97。当用 MAE 作为预测精度的测度时，线性趋势回归同样提供更准确的预测：对自行车销售量时间序列，线性趋势回归的 MAE = 1.32，而用 Holt 线性方法的 MAE = 1.67。然而，基于 MAPE，Holt 线性指数平滑（MAPE = 5.07%）的表现优于线性趋势回归（MAPE = 6.42%）。因此，对自行车销售量时间序列，哪一种方法提供更准确的预测取决于使用哪一种预测精度的测度。



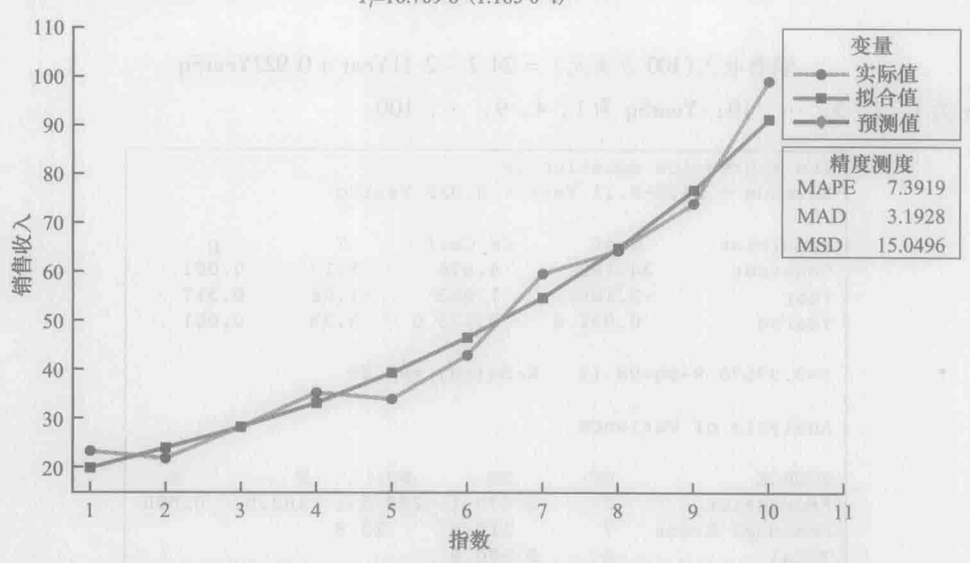


图 17-15 胆固醇销售收入时间序列的 Minitab 时间序列指数增长趋势分析的输出结果

练 习

方法

18. 参见第 17 题的时间序列数据。用 $\alpha=0.3$ 和 $\beta=0.5$ 的 Holt 线性指数平滑方法计算 $t=6$ 的预测值。

20. 考虑下面的时间序列。

t	1	2	3	4	5	6	7
Y_t	84	60	44	35	30	29	35

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 利用 Minitab 或 Excel，建立这个时间序列的二次趋势方程。

c. $t=8$ 的预测值是多少？

22. Seneca 儿童基金会（SCF）是一家为贫困儿童举办夏令营的当地慈善机构。近年来基金会主席致力于减少经常性支出的总额，主要因素是慈善机构被认为是独立机构。下面的数据表示用于管理和筹集资金的费用占 SCF 募集到的总资金的比例。

年	时期（t）	费用（%）
2003	1	13.9
2004	2	12.2
2005	3	10.5
2006	4	10.4
2007	5	11.5
2008	6	10.0
2009	7	8.5

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 建立这个时间序列的线性趋势方程。

c. 预测 2010 年管理费用的比例。

d. 如果 SCF 保持当前减少管理费用的趋势，要达到小于或等于 5% 的水平需要多长时间？

24. 联邦储备经济数据（Federal Reserve Economic Data，FRED®）是一个有 3 000 多个美国经济时间序列的数据库，包含有外汇汇率的历史数据。下面是美国和中国的外汇汇率数据（Federal Reserve Bank of St. Louis website），汇率单位是 1 美元兑换人民币元的数量。

年	月	汇率
2007	10	7.501 9
2007	11	7.421 0
2007	12	7.368 2
2008	1	7.240 5
2008	2	7.164 4
2008	3	7.072 2
2008	4	6.999 7
2008	5	6.972 5
2008	6	6.899 3
2008	7	6.835 5

a. 绘制时间序列的图，是否存在线性趋势？

b. 利用 Minitab 或 Excel，建立这个时间序列的线

性趋势方程。

c. 利用趋势方程预测2008年8月的汇率。

d. 利用趋势方程预测2008年12月的汇率，你认为合适吗?

26. Giovanni食品生产公司为遍及美国东部的公立中小学校生产和销售冷冻比萨饼。利用一个非常积极的营销战略，在过去10年间，他们的年收入已经增加了近1 000万美元。日益激烈的竞争使得他们的增长速度在过去几年已经放缓。10年期间的年收入（单位：100万美元）如下。

年	收入	年	收入
1	8.53	6	17.21
2	10.84	7	18.37
3	12.98	8	18.45
4	14.11	9	18.40
5	16.31	10	18.43

a. 绘制时间序列的散点图，评价线性趋势的合理性。

b. 利用Minitab或Excel，建立用于预测销售量的二次趋势方程。

c. 利用（b）中建立的趋势方程，预测第11年的收入。

17.5 季节性和趋势

本节我们演示如何对有季节模式的时间序列建立预测。在存在季节性的情形下，我们需要将其并入我们的预测模型以确保预测的准确性。我们首先考虑没有趋势的季节性时间序列，然后讨论如何模拟有趋势的季节性。

17.5.1 没有趋势的季节性

作为一个例子，考虑过去5年某服装商店的雨伞销售量。表17-17是时间序列，图17-16是相应的时间序列图。时间序列图没有显示销售量有长期趋势。事实上，如果不仔细观测数据，你可能会得出数据具有水平模式的结论，并用单一指数平滑法预测销售量。但是，仔细检查时间序列图，发现数据存在模式，即第一季度和第三季度有中等销售量，第二季度的销售量最高，而第四季度往往是最低销售量的季度。因此，我们得出存在季度季节模式的结论。

表17-17 雨伞销售量的时间序列

年份	季度	销售量	年份	季度	销售量
1	1	125	3	3	113
	2	153		4	80
	3	106	4	1	109
	4	88		2	137
2	1	118		3	125
	2	161		4	109
	3	133	5	1	130
	4	102		2	165
3	1	138		3	128
	2	177		4	96

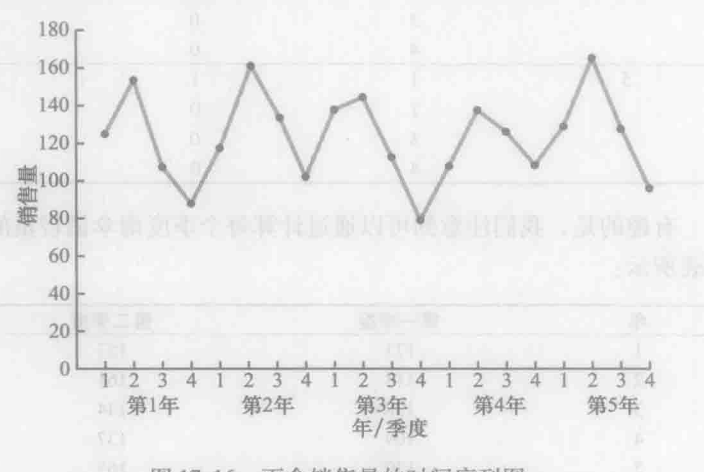


图17-16 雨伞销售量的时间序列图

在第15章我们介绍了在多元回归模型中，如何用虚拟变量处理分类自变量。我们可以将季节作为一个分类变量，使用同样的方法去模拟具有季节模式的时间序列。回忆一下，当分类变量有 k 个水平时，需要 $k-1$ 个虚拟变量。因此，如果有4个季度，我们需要3个虚拟变量。例如，对雨伞销售量时间序列，季度是一个有4个水平的分类变量：第一季度、第二季度、第三季度和第四季度。因此，为了模拟雨伞时间序列的季节影响，我们需要 $4-1=3$ 个虚拟变量。将3个虚拟变量编码如下：

$$\text{Qtr1}=\begin{cases}1, & \text{第一季度}\\0, & \text{其他}\end{cases}\qquad \text{Qtr2}=\begin{cases}1, & \text{第二季度}\\0, & \text{其他}\end{cases}\qquad \text{Qtr3}=\begin{cases}1, & \text{第三季度}\\0, & \text{其他}\end{cases}$$

用 $\hat{y}$ 表示销售量的估计值或预测值，与雨伞季度销售量关于的估计的回归方程的一般形式如下：

$$\hat{Y} = b_0 + b_1\text{Qtr1} + b_2\text{Qtr2} + b_3\text{Qtr3}$$

表 17-18 是雨伞销售量时间序列以及虚拟变量的编码值。用表 17-18 中的数据以及 Minitab 回归程序，我们得到图 17-17 的计算机输出结果。得到的估计的多元回归方程是

$$\text{Sales} = 95.0 + 29.0\text{Qtr1} + 57.0\text{Qtr2} + 26.0\text{Qtr3}$$

我们能利用这个方差预测下一年的季度销售量。

$$第一季度：\text{Sales} = 95.0 + 29.0 \times 1 + 57.0 \times 0 + 26.0 \times 0 = 124$$

$$第二季度：\text{Sales} = 95.0 + 29.0 \times 0 + 57.0 \times 1 + 26.0 \times 0 = 152$$

$$第三季度：\text{Sales} = 95.0 + 29.0 \times 0 + 57.0 \times 0 + 26.0 \times 1 = 121$$

$$第四季度：\text{Sales} = 95.0 + 29.0 \times 0 + 57.0 \times 0 + 26.0 \times 0 = 95$$

表 17-18　具有虚拟变量的雨伞销售量时间序列

年	季度	Qtr1	Qtr2	Qtr3	销售量
1	1	1	0	0	125
	2	0	1	0	153
	3	0	0	1	106
	4	0	0	0	88
2	1	1	0	0	118
	2	0	1	0	161
	3	0	0	1	133
	4	0	0	0	102
3	1	1	0	0	138
	2	0	1	0	144
	3	0	0	1	113
	4	0	0	0	80
4	1	1	0	0	109
	2	0	1	0	137
	3	0	0	1	125
	4	0	0	0	109
5	1	1	0	0	130
	2	0	1	0	165
	3	0	0	1	128
	4	0	0	0	96

有趣的是，我们注意到可以通过计算每个季度雨伞销售量的简单平均数就可以得到下一年季度的预测值，如下表所示：

年	第一季度	第二季度	第三季度	第四季度
1	125	153	106	88
2	118	161	133	102
3	138	144	113	80
4	109	137	125	109
5	130	165	128	96
平均数	124	152	121	95

尽管如此，图 17-17 中的回归输出结果还给出了能评价预测准确性和确定结果显著性的附加信息。同时，对问题情况更复杂的类型，如涉及趋势和季节影响同时存在的时间序列，简单平均方法将无效。

```
The regression equation is
Sales=95.0+29.0  Qtr1+57.0  Qtr2+26.0  Qtr3

Predictor     Coef      SE Coef      T       P
Constant     95.000      5.065     18.76   0.000
Qtr1         29.000      7.162      4.05   0.001
Qtr2         57.000      7.162      7.96   0.000
Qtr3         26.000      7.162      3.63   0.002
```

图 17-17　雨伞销售量时间序列的 Minitab 回归输出结果

17.5.2　季节性和趋势

现在让我们以第 17.1 节介绍的电视机销售量时间序列

为例，将回归方法扩展到季节影响和线性趋势同时存在的时间序列情形，并演示如何进行季度预测。表 17-19 是电视机销售量时间序列的数据，图 17-18 的时间序列图显示在每年第二季度销售量最低，第三和第四季度销售量增长。因此，我们得出电视机销售量存在季节模式。但是时间序列还有为了建立季度销售量的精确预测需要解释的一个上升的线性趋势。这是很容易处理的，只要将季节性的虚拟变量的方法与我们在第 17.3 节讨论的处理线性趋势的时间序列回归方法结合起来就行。

模拟电视机时间序列的季度季节影响和线性趋势的多元回归方程的一般形式如下：

$$\hat{Y}_t = b_0 + b_1\text{Qtr1} + b_2\text{Qtr2} + b_3\text{Qtr3} + b_4 t$$

式中，$\hat{Y}_t$ 为 t 期销售量的估计值或预测值；Qtr1 = 1，如果对应于 t 期年份的第一季度，否则 Qtr1 = 0；Qtr2 = 1，如果对应于 t 期年份的第二季度，否则 Qtr2 = 0；Qtr3 = 1，如果对应于 t 期年份的三季度，否则 Qtr3 = 0；t 为时期。



表 17-19　电视机销售量的时间序列

年份	季度	销售量（1 000 台）	年份	季度	销售量（1 000 台）
1	1	4.8	3	1	6.0
	2	4.1		2	5.6
	3	6.0		3	7.5
	4	6.5		4	7.8
2	1	5.8	4	1	6.3
	2	5.2		2	5.9
	3	6.8		3	8.0
	4	7.4		4	8.4

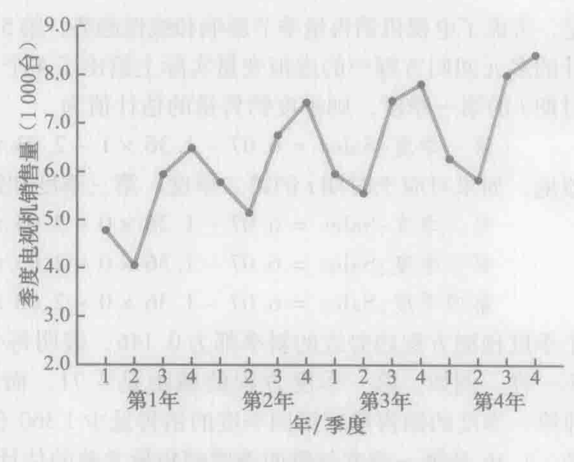


图 17-18　电视机销售量的时间序列图

表 17-20 是电视机销售量时间序列，包含虚拟变量的编码值和时期 t。用表 17-20 中的数据以及 Minitab 回归程序，我们得到图 17-19 的计算机输出结果。估计的多元回归方程为

$$\text{Sales} = 6.07 - 1.36\text{Qtr1} - 2.03\text{Qtr2} - 0.304\text{Qtr3} + 0.146t \qquad (17\text{-}12)$$

表 17-20　具有虚拟变量和时期的电视机销售量时间序列

年	季度	Qtr1	Qtr2	Qtr3	时期	销售量（1 000 台）
1	1	1	0	0	1	4.8
	2	0	1	0	2	4.1
	3	0	0	1	3	6.0
	4	0	0	0	4	6.5
2	1	1	0	0	5	5.8
	2	0	1	0	6	5.2
	3	0	0	1	7	6.8
	4	0	0	0	8	7.4
3	1	1	0	0	9	6.0
	2	0	1	0	10	5.6
	3	0	0	1	11	7.5
	4	0	0	0	12	7.8
4	1	1	0	0	13	6.3
	2	0	1	0	14	5.9
	3	0	0	1	15	8.0
	4	0	0	0	16	8.4

现在我们用式（17-12）来预测下一年销售量的季度预测值。下一年是电视机销售量时间序列的第 5 年，即

时期为 17，18，19 和 20。

时期 17（第 5 年第一季度）的预测值为

$$Sales = 6.07 - 1.36 \times 1 - 2.03 \times 0 - 0.304 \times 0 + 0.146 \times 17 = 7.19$$

时期 18（第 5 年第二季度）的预测值为

$$Sales = 6.07 - 1.36 \times 0 - 2.03 \times 1 - 0.304 \times 0 + 0.146 \times 18 = 6.67$$

```
The regression equation is
Sales (1 000s)=6.07-1.36  Qtr1-2.03  Qtr2-0.304  Qtr3 + 0.146 Period

Predictor    Coef       SE Coef     T       P
Constant     6.068 8    0.162 5     37.35   0.000
Qtr1        -1.363 1    0.157 5     -8.66   0.000
Qtr2        -2.033 7    0.155 1    -13.11   0.000
Qtr3        -0.304 4    0.153 7     -1.98   0.073
Period       0.1456 2   0.012 11    12.02   0.000
```

图 17-19 电视机销售量时间序列的 Minitab 回归输出结果

时期 19（第 5 年第三季度）的预测值为

$$Sales = 6.07 - 1.36 \times 0 - 2.03 \times 0 - 0.304 \times 1 + 0.146 \times 19 = 8.54$$

时期 20（第 5 年第四季度）的预测值为

$$Sales = 6.07 - 1.36 \times 0 - 2.03 \times 0 - 0.304 \times 0 + 0.146 \times 20 = 8.99$$

于是，考虑了电视机销售量季节影响和线性趋势，第 5 年季度销售量的估计值为 7 190，6 670，8 540 和 8 990。

估计的多元回归方程中的虚拟变量实际上给出了 4 个（每个季度各 1 个）估计的多元回归方程，例如，如果对应于时期 t 的第一季度，则季度销售量的估计值为

$$\text{第一季度}: Sales = 6.07 - 1.36 \times 1 - 2.03 \times 0 - 0.304 \times 0 + 0.146t = 4.71 + 0.146t$$

类似地，如果对应于时期 t 的第二季度、第三季度和第四季度，则季度销售量的估计值为

$$\text{第二季度}: Sales = 6.07 - 1.36 \times 0 - 2.03 \times 1 - 0.304 \times 0 + 0.146t = 4.04 + 0.146t$$

$$\text{第三季度}: Sales = 6.07 - 1.36 \times 0 - 2.03 \times 0 - 0.304 \times 1 + 0.146t = 5.77 + 0.146t$$

$$\text{第四季度}: Sales = 6.07 - 1.36 \times 0 - 2.03 \times 0 - 0.304 \times 0 + 0.146t = 6.07 + 0.146t$$

每个季度预测方程趋势线的斜率都为 0.146，表明每个季度销售量大约增长 146 台。唯一不同的是 4 个方程的截距不一样。例如，第一季度方程的截距是 4.71，而第四季度方程的截距是 6.07。因此，4.71 - 6.07 = -1.36，即第一季度的销售量较第四季度的销售量少 1 360 台。换句话说，方程（17-12）中虚拟变量 Qtr1 的估计的回归系数 -1.36 是第一季度与第四季度销售量之差的估计值。类似地，可以解释虚拟变量 Qtr2 的估计的回归系数 -2.03，以及虚拟变量 Qtr3 的估计的回归系数 -0.304。

17.5.3 基于月度数据的模型

在前面电视机销售量的例子中，我们演示了如何用虚拟变量来说明时间序列的季度季节影响。由于分类变量季度有 4 个水平，需要 3 个虚拟变量[⊖]。然而，许多商务问题用月度而不是季度进行预测。对月度数据，季节是一个有 12 个水平的分类变量，因此需要 12 - 1 =11 个虚拟变量。例如，11 个虚拟变量编码如下：

$$Month1 = \begin{cases} 1, & 1\text{月} \\ 0, & \text{其他} \end{cases}, Month2 = \begin{cases} 1, & 2\text{月} \\ 0, & \text{其他} \end{cases}, \cdots, Month11 = \begin{cases} 1, & 11\text{月} \\ 0, & \text{其他} \end{cases}$$

除了这些变化，处理季节性的多元回归方法是一样的。

练 习

方法

☆28. 考虑下面的时间序列：

季度	年份 1	年份 2	年份 3
1	71	68	62
2	49	41	51
3	58	60	53
4	78	81	72

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 用下面的虚拟变量来建立一个解释数据中季节影响的估计的趋势方程：如果为第一季度，Qtr1 = 1，否则为 0；如果为第二季度，Qtr2 = 1，否则为 0；如果为第三季度，Qtr3 = 1，否则为 0。

c. 计算下一年的季度预测值。

⊖ 当一个分类变量（如季节）有 k 个水平时，需要 $k-1$ 个虚拟变量。

应用

30. 过去三年中大学教材的季度销售量（销售的册数）数据如下：

季度	年份 1	年份 2	年份 3
1	1 690	1 800	1 850
2	940	900	1 100
3	2 625	2 900	2 930
4	2 500	2 360	2 615

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 用下面虚拟变量来建立一个解释数据中季节影响的估计的趋势方程：如果为第一季度，Qtr1 = 1，否则为 0；如果为第二季度，Qtr2 = 1，否则为 0；如果为第三季度，Qtr3 = 1，否则为 0。

c. 计算下一年的季度预测值。

d. 令 $t=1$ 表示第 1 年第一季度的观测值，$t=2$ 表示第 1 年第二季度的观测值，…，$t=12$ 表示第 3 年第四季度的观测值。用（b）定义的虚拟变量和 t，建立一个解释时间序列中季节影响和线性趋势估计的回归方程。基于数据中的季节影响和线性趋势，计算下一年的预测值。

32. South Share 建筑公司沿着纽约长岛南岸建立永久性码头和防波堤。尽管公司开业至今仅有 5 年，但收入却从经营第 1 年的 308 000 美元增长到最近一年的 1 084 000 美元。下面是季度收入数据（单位：1 000 美元）。

季度	年份 1	年份 2	年份 3	年份 4	年份 5
1	20	37	75	92	176
2	100	136	155	202	282
3	175	245	326	384	445
4	13	48	48	82	181

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 用下面虚拟变量建立一个解释数据中季节影响的估计的趋势方程：如果为第一季度，Qtr1 = 1，否则为 0；如果为第二季度，Qtr2 = 1，否则为 0；如果为第三季度，Qtr3 = 1，否则为 0。基于数据中的季节影响，计算第 6 年的季度销售收入的预测值。

c. 令 Period = 1 表示第 1 年第一季度的观测值，Period = 2 表示第 1 年第二季度的观测值，…，Period = 20 表示第 5 年第四季度的观测值。用（b）中定义的虚拟变量和 Period，建立一个解释时间序列中季节影响和线性趋势估计的回归方程。基于数据中的季节影响和线性趋势，计算第 6 年的季度销售收入的预测值。

34. 下面是佛罗里达州南部一个 6 单元公寓 3 年的月草坪维护费用（单位：美元）。

月份	年份 1	年份 2	年份 3
1	170	180	195
2	180	205	210
3	205	215	230
4	230	245	280
5	240	265	290
6	315	330	390
7	360	400	420
8	290	335	330
9	240	260	290
10	240	270	295
11	230	255	280
12	195	220	250

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 建立一个用于解释数据中季节和线性趋势影响的估计的回归方程。用下面虚拟变量来解释数据中的季节影响：如果为 1 月，Jan = 1，否则为 0；如果为 2 月，Feb = 1，否则为 0；如果为 3 月，Mar = 1，否则为 0；…，如果为 11 月，Nov = 1，否则为 0。注意，用这个编码方法，当 11 个虚拟变量的值都为 0 时，对应于 12 月费用的观测值。

c. 基于趋势和季节双重影响计算下一年的月度预测值。

17.6 时间序列分解法

本节我们关注**时间序列分解**（time series decomposition）的问题。用时间序列分解法可以将一个时间序列分隔或分解出季节、趋势和不规则成分。当用这种方法进行预测时，其主要应用是获得对时间序列的一个更好的了解。许多商务和经济时间序列由政府机构（如人口普查局和劳工统计局）保存和出版，这些机构用时间序列分解来建立消除季节影响的时间序列。

了解时间序列的真实状况常常依赖于使用消除季节影响的数据。例如，我们或许有兴趣了解我们地区的电力消耗是否正在增长。假设我们了解到，9 月份的电力消费量比 8 月份减少了 3%。使用这些资料时必须相当谨慎，

因为每当存在季节影响时，如果数据没有消除季节影响，这样的比较可能会使人误解。9 月份的电力消费量比 8 月份减少了 3% 的事实，可能仅仅是与空调使用减少有关的季节影响，而并不是因为用电量的长期减少。事实上，在调整季节影响之后，我们甚至可能发现用电量是增加的。许多其他时间序列，如失业率统计、房屋销售量和零售额，都受到很强的季节影响。在对这些数据作出长期趋势的判断之前，消除季节影响是非常重要的。

时间序列分解方法假设时间序列 t 期的实际值 Y_t，是趋势成分、季节成分和不规则或误差成分这三个成分的函数。如何将这三个成分组合起来产生时间序列的观测值，取决于我们假定是用加法模型还是用乘法模型来更好地描述它们之间的关系。

加法分解模型（additive decomposition model）的形式如下：

$$Y_t = \text{Trend}_t + \text{Seasonal}_t + \text{Irregular}_t \tag{17-13}$$

式中，Trend_t 为 t 期的趋势值；Seasonal_t 为 t 期的季节值；Irregular_t 为 t 期的不规则值。

在加法模型中，三个成分的值简单加在一起得到时间序列的实际值 Y_t，不规则或误差成分[㊀]说明时间序列中不能被趋势和季节成分所解释的变异性。

加法模型适合于季节影响不依赖于时间序列水平的情形。第 17.5 节包含季节和趋势影响的回归模型是一个加法模型。如果前期季节影响的规模与后期季节影响的规模相同，则加法模型是合适的。然而，如果季节影响随时间发生改变，由于长期线性趋势随着销售量的增加逐渐变大，则应该使用乘法模型[㊁]。许多商务与经济时间序列都遵循这种模式。

乘法分解模型（multiplicative decomposition model）的形式如下：

$$Y_t = \text{Trend}_t \times \text{Seasonal}_t \times \text{Irregula}_t \tag{17-14}$$

式中，Trend_t 为 t 期的趋势值；Seasonal_t 为 t 期的季节值；Irregular_t 为 t 期的不规则值。

在这个模型中，趋势、季节和不规则成分相乘可以得到时间序列的值。趋势用被预测项目的单位来度量，但是，季节和不规则成分则按相对量来度量，其数值大于 1.00 表明它们的影响在趋势之上，而数值小于 1.00 表明它们的影响在趋势之下。

由于这种方法在实践中常常使用，因此，介绍如何建立趋势和季节成分的估计而对时间序列分解的讨论，我们将只限定于乘法模型。我们以第 17.5 节介绍的电视机季度销售量时间序列为例进行说明；表 17-19 中是季度销售量数据，图 17-18 是相对应的时间序列图。在演示如何用乘法模型分解时间序列后，我们将说明如何将季节指数和趋势成分重新组合起来进行预测。

17.6.1 计算季节指数

图 17-18 显示每年的第二季度销售量最小，第三和第四季度的销售量开始增加。因此，我们推断电视机销售量时间序列存在季节模式。确定每个季度的季节影响所使用的计算程序是先计算移动平均数，从数据中剔除组合在一起的季节和不规则影响，留给我们的时间序列只包含趋势和移动平均没有剔除的随机波动。

由于我们所使用的数据是季度时间序列，因此在每一次移动平均数的计算中，我们使用 4 项数据。前 4 个季度的电视机销售量的移动平均数是

$$\text{第一个移动平均数} = \frac{4.8 + 4.1 + 6.0 + 6.5}{4} = \frac{21.4}{4} = 5.35$$

注意，前 4 个季度的移动平均数正好是时间序列第 1 年的平均季度销售量。继续移动平均数的计算，加上第 2 年第一季度的数值 5.8，去掉第 1 年第一季度的数值 4.8，得到第二个移动平均数为

$$\text{第二个移动平均数} = \frac{4.1 + 6.0 + 6.5 + 5.8}{4} = \frac{22.4}{4} = 5.60$$

同理，第三个移动平均数为 $(6.0 + 6.5 + 5.8 + 5.2)/4 = 5.875$。

㊀ 不规则成分与我们在第 14 章中讨论的简单线性回归模型中的误差项 ε 相对应。

㊁ 人口普查局使用了一个与该局消除季节影响后的时间序列相结合的乘法模型。

在我们对整个时间序列进行移动平均数计算之前，我们再回到第一个移动平均数的计算，得到5.35。5.35这个数字正好是第1年的平均季度销售量。当我们回顾5.35的计算结果时，发现它对应于移动平均数的所有季度的“中间”。但是，请注意，因为是4个季度的移动平均数，不存在中间季度，因此，5.35应该对应于2.5季度，即第二季度的后半部分和第三季度的前半部分。同样，如果我们考察下一个移动平均数为5.60，它对应于的中间季度为3.5，即第三季度的后半部分和第四季度的前半部分。

我们计算出的两个移动平均数并不能直接对应于时间序列的原始季度，因此，我们可以通过计算两个移动平均数的平均数来解决这个难题。既然第一个移动平均数的中心是2.5季度（季度的一半或前半个季度），第二个移动平均数的中心是3.5季度（季度的一半或后半个季度），则两个移动平均数的平均数的中心为3季度，这应该是它精确的位置。这个移动平均数被称为中心化移动平均数。因此，第三季度的中心化移动平均数为（5.35 + 5.60）/2 = 5.475，同样，第四季度的中心化移动平均数为（5.60 + 5.875）/2 = 5.738。表17-21是电视机销售量数据的移动平均数和中心化移动平均数的全面汇总。

表 17-21　电视机销售量时间序列的中心化移动平均数的计算

年	季度	销售量（1 000 件）	4 个季度的移动平均数	中心化的移动平均数
1	1	4.8		
	2	4.1		
			5.350	
	3	6.0		5.475
			5.600	
	4	6.5		5.738
			5.875	
2	1	5.8		5.975
			6.075	
	2	5.2		6.188
			6.300	
	3	6.8		6.325
			6.350	
	4	7.4		6.400
			6.450	
3	1	6.0		6.538
			6.625	
	2	5.6		6.675
			6.725	
	3	7.5		6.763
			6.800	
	4	7.8		6.838
			6.875	
4	1	6.3		6.938
			7.000	
	2	5.9		7.075
			7.150	
	3	8.0		
	4	8.4		

表17-21的中心化的移动平均数能告诉我们时间序列的什么问题呢？图17-20是时间序列实际值和中心化的移动平均数的时间序列图。要特别注意，中心化的移动平均数是如何“平滑”时间序列的季节和不规则波动影响的。中心化的移动平均数描绘数据中的趋势和没有被平滑数据的移动平均剔除掉的随机波动。

前面我们介绍的乘法分解模型为

$$Y_t = \text{Trend}_t \times \text{Seasonal}_t \times \text{Irregular}_t$$

方程两边同时除以趋势成分 T_t，我们可以识别出时间序列中的季节－不规则[㊀]的组合影响为

㊀ 季节－不规则值常常被称为时间序列的被剔除趋势的数值。

$$\frac{Y_t}{\text{Trend}_t} = \frac{\text{Trend}_t \times \text{Seasonal}_t \times \text{Irregular}_t}{\text{Trend}_t}$$
$$= \text{Seasonal}_t \times \text{Irregular}_t$$

例如，第1年第三季度的趋势值为5.475（中心化移动平均数），因此，6.0/5.475 = 1.096是季节－不规则的组合值。表17-22汇总了整个时间序列的季节－不规则值。

表17-22 电视机销售量时间序列的季节－不规则值

年	季度	销售量（1 000台）	中心化移动平均数	季节－不规则值
1	1	4.8		
	2	4.1		
	3	6.0	5.475	1.096
	4	6.5	5.738	1.133
2	1	5.8	5.975	0.971
	2	5.2	6.188	0.840
	3	6.8	6.325	1.075
	4	7.4	6.400	1.156
3	1	6.0	6.538	0.918
	2	5.6	6.675	0.839
	3	7.5	6.763	1.109
	4	7.8	6.838	1.141
4	1	6.3	6.938	0.908
	2	5.9	7.075	0.834
	3	8.0		
	4	8.4		

下面考虑第三季度的季节－不规则值：1.096，1.075和1.109。季节－不规则值大于1.00表明影响在趋势估计之上，而数值小于1.00表明影响在趋势估计之下。因此，在第三季度，3个季节－不规则值表现出高于平均水平的影响。因为季节－不规则值每年的波动主要是由于随机误差引起的，所以，我们可以取其平均数以消除不规则的影响，从而得到第三季度季节影响的估计值。

$$\text{第三季度的季节影响值} = \frac{1.096 + 1.075 + 1.109}{3} = 1.09$$

我们将1.09称为第三季度的季节指数。在表17-23中，汇总了关于电视机销售量时间序列的季节指数的计算结果。4个季度的季节指数分别为：0.93，0.84，1.09和1.14。

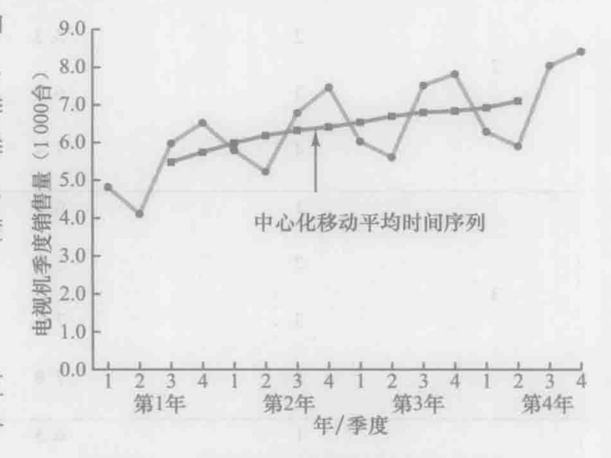


图17-20 电视机季度销售量时间序列和中心化移动平均数

表17-23中季节指数的解释提供了关于电视机销售量季节成分的一些观测结果。最佳销售季度是第四季度，其销售水平高出趋势估计14%，最差或最少的销售季度是第二季度，其季节指数为0.84，表明其销售水平低于趋势估计16%。季节成分很明显地对应着人们观看电视兴趣的直观期望，由于冬季即将到来人们减少了户外活动，因此电视购买模式趋向于峰值。第二季度的低销售量反映由于春季和初夏的户外活动，减少了潜在消费者观看电视的兴趣。

表17-23 电视机销售量时间序列的季节指数的计算

季度	季节－不规则成分的数值			季节指数
1	0.971	0.918	0.908	0.93
2	0.840	0.839	0.934	0.84
3	1.096	1.075	1.109	1.09
4	1.113	1.156	1.141	1.14

在取得季节指数时，对季节指数做最后的调整有时是必要的。乘法模型要求平均季节指数等于1.00，因此，表17-23中的4个季节指数之和必须等于4.00。换句话说，季节影响在一年内必须是持平的。在我们这个例子中季节指数的平均数等于1.00，因此，它不需要调整。对于其他情形，有时或许需要进行微调。对季节指数进行调整是用每个季节指数乘以季度总和再除以未调整的季节指数之和。例如，对于季度数据，每个季节指数乘以4再除以未调整季节指数之和。为了获得合理的季节指数，有些练习需要进行这种调整。

17.6.2 消除季节影响的时间序列

剔除了季节影响的时间序列被称为**消除季节影响的时间序列**（deseasonalized time series），用季节指数剔除时间序列中的季节影响的过程称为消除时间序列的季节影响。[⊖]利用乘法分解模型，我们用每一个观测值除以相应的季节指数来消除时间序列的季节影响。乘法分解模型为

$$Y_t = \text{Trend}_t \times \text{Seasonal}_t \times \text{Irregular}_t$$

因此，当时间序列的每个观察值 Y_t 除以相应的季节指数时，剩下的数据中只包含趋势和随机波动（不规则成分）。在表17-24中汇总了消除季节影响后的电视机销售量的时间序列，图17-21是消除季节影响的时间序列的图形。

表17-24 消除季节影响的电视机销售量的时间序列值

年	季度	时期	销售量（1 000台）	季节指数	消除季节影响的销售量
1	1	1	4.8	0.93	5.16
	2	2	4.1	0.84	4.88
	3	3	6.0	1.09	5.50
	4	4	6.5	1.14	5.70
2	1	5	5.8	0.93	6.24
	2	6	5.2	0.84	6.19
	3	7	6.8	1.09	6.24
	4	8	7.4	1.14	6.49
3	1	9	6.0	0.93	6.45
	2	10	5.6	0.84	6.67
	3	11	7.5	1.09	6.88
	4	12	7.8	1.14	6.84
4	1	13	6.3	0.93	6.77
	2	14	5.9	0.84	7.02
	3	15	8.0	1.09	7.34
	4	16	8.4	1.14	7.37

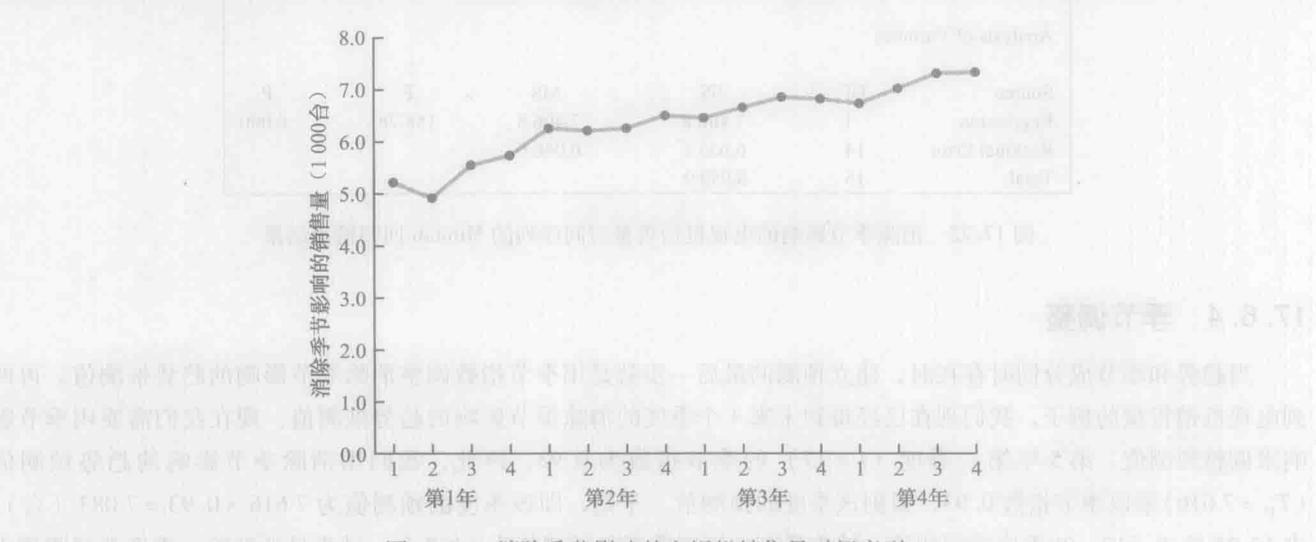


图17-21 消除季节影响的电视机销售量时间序列

⊖ 调整季节变异后的经济时间序列常常刊登在一些出版物上，如《当代商业纵览》《华尔街日报》和《商业周刊》。

17.6.3 利用消除季节影响的时间序列确定趋势

出现在图 17-21 中的消除季节影响的电视机销售量时间序列的图形，显示有一个向上的线性趋势。为了确定这个趋势，我们使用第 17.4 节介绍的同样方法，用一个线性趋势方程来拟合消除季节影响的时间序列。唯一不同的是，我们使用消除季节影响的数据而不是原始数据。

我们记得，对线性趋势，估计的回归方程可以表示为

$$T_t = b_0 + b_1 t$$

式中，Y_t 是 t 期线性趋势的预测值；b_0 是线性趋势线的截距；b_1 是趋势线的斜率；t 是时期。

在第 17.4 节我们给出了计算 b_0 和 b_1 的公式。为了用一条线性趋势线拟合表 17-24 的消除季节影响的数据，在计算 b_0 和 b_1 时，唯一的变化是我们用消除季节影响的时间序列值代替观测值 Y_t。

图 17-22 是利用 Minitab 回归分析程序来估计消除季节影响的电视机销售量时间序列趋势线的计算机输出结果。估计的线性趋势方程为

$$\text{消除季节影响的销售量} = 5.10 + 0.148t$$

斜率 0.148 表明，在过去的 16 个季度中，消除季节影响之后，公司每个季度的销售量平均增长 148 台。如果我们假设，过去 16 个季度销售的趋势依然适用于未来，则这个方程可用来建立未来季度的趋势推测。例如，将 $t=17$代入方程，可以得到下一个季度消除季节影响的趋势推测值 T_{17}。

$$T_{17} = 5.10 + 0.148t = 0.51 + 0.148 \times 17 = 7.616$$

因此，利用消除季节影响的数据，下一个季度（时期 17）的线性趋势预测值 7 616 台电视机。类似地，下 3 个季度（时期 18，19 和 20）的消除季节影响的趋势预测值分别为 7 764，7 912 和 8 060 台电视机。

```
The regression equation is
Deseasonalized Sales=5.10+0.148 Period

Predictor       Coef       SE Coef     T       P
Constant        5.105 0    0.113 3     45.07   0.000
Period          0.147 60   0.011 71    12.60   0.000

S=0.215 985     R-Sq=91.9%        R-Sq(adj)=91.3%

Analysis of Variance

Source          DF    SS         MS         F        P
Regression      1     7.406 8    7.406 8    158.78   0.000
Residual Error  14    0.653 1    0.046 6
Total           15    8.059 9
```

图 17-22 消除季节影响的电视机销售量时间序列的 Minitab 回归输出结果

17.6.4 季节调整

当趋势和季节成分同时存在时，建立预测的最后一步就是用季节指数调整消除季节影响的趋势推测值。再回到电视机销售量的例子，我们现在已经得到未来 4 个季度的消除季节影响的趋势推测值，现在我们需要用季节影响来调整预测值。第 5 年第一季度（$t=17$）的季节指数为 0.93，因此，我们用消除季节影响的趋势预测值（$T_{17}=7\,616$）乘以季节指数 0.93，得到该季度的预测值。于是，即该季度的预测值为 7 616 × 0.93 = 7 083（台）。表 17-25 给出了17 ~ 20季度的预测值。销售最高的第四季度的预测值为 9 188 台；销售最低的第二季度的预测值为 6 522 台。

表 17-25　电视机销售量时间序列的季度预测值

年	季度	消除季节影响的趋势预测值	季节指数	季度预测值
5	1	7 616	0.93	7 616 × 0.93 = 7 083
	2	7 764	0.84	7 764 × 0.84 = 6 522
	3	7 912	1.09	7 912 × 1.09 = 8 624
	4	8 060	1.14	8 060 × 1.14 = 9 188

17.6.5　根据月度数据的模型

在前面电视机销售量的例子中，我们利用季度数据来说明季节指数的计算。然而，许多商务问题使用的是月度而不是季度的预测。在这种情况下，本节介绍的方法只要做轻微的修改就可以使用。首先，是用 12 个月的移动平均数代替 4 个季度的移动平均数；其次，是计算 12 个月的季节指数，而不是 4 个季度的季节指数。除此之外，计算和预测方法都是一样的。

17.6.6　循环成分

在数学上，式（17-14）的乘法模型可以扩展到包括循环成分。

$$Y_t = \text{Trend}_t \times \text{Cyclical}_t \times \text{Seasonal}_t \times \text{Irregular}_t \tag{17-15}$$

循环成分同季节成分一样，可以表示为趋势的百分比。正如第 17.1 节所述，循环成分归因于时间序列的多年循环。它与季节成分类似，只是所经历的时间周期更长一些。但是，由于所涉及的时间的长度，要获得足够的相关数据来估计循环成分常常是困难的。另一个困难是循环的周期长短不一。由于从长期趋势影响中确定和（或）分离循环影响如此困难，实践中，常常将这些影响合并，称之为趋势 - 循环组合成分。我们不再对循环成分做进一步的讨论，在预测方法的专业教材中有这方面的内容。

注释和评论

1. 有许多不同的计算季节指数的方法。在本节我们通过平均对应的季节 - 不规则值计算每一个季节指数。另一种被 Minitab 所使用的方法是用季节 - 不规则值的中位数作为季节指数。
2. 在消除时间序列的季节影响之前，常常需要做日历调整。例如，如果时间序列是月销售量数值，则 2 月份的销售量数值可能比其他月份的数值小，这仅仅是因为 2 月份的天数少。为了避免这个因素，我们首先用每个月的销售数值除以该月的天数，得到日销售量的平均数。因为每个月平均大约有 365/12 = 30.416 7（天），然后我们用日销售量的平均数乘以 30.416 7，得到调整的月销售量数值。对于本章的例题和练习，你可以认为已经进行了必要的日历调整。

练　习

方法

36. 参考第 35 题。
    a. 用第 35 题（c）中的调整季节指数来消除时间序列的季节影响。
    b. 用 Minitab 或 Excel，计算消除季节影响数据的线性趋势回归方程。
    c. 计算第 4 年的消除季节影响的季度趋势预测值。
    d. 用季节指数调整（c）中的消除季节影响的季度趋势预测值。

应用

38. 下面是佛罗里达州南部一个 6 单元的公寓在 3 个年份的月草坪维护费用（单位：美元）。

月份	年份 1	年份 2	年份 3
1	170	180	195
2	180	205	210
3	205	215	230
4	230	245	280
5	240	265	290
6	315	330	390
7	360	400	420

（续）

月份	年份1	年份2	年份3
8	290	335	330
9	240	260	290
10	240	270	295
11	230	255	280
12	195	220	250

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 确定佛罗里达州南部公寓3年的月草坪维护费用的每月的季节指数。用12个月的移动平均数计算。

c. 消除时间序列的季节影响。

d. 计算消除季节影响数据的线性趋势方程。

e. 计算消除季节影响的趋势预测值，然后用季节指数调整趋势预测值，从而得到第4年每月费用的预测值。

40. 电能的消耗量用千瓦时（kWh）来度量。某地公用事业公司主动提供一个需要商业客户参与的电力供应中断程序。如果客户参与该程序，则电费将会得到优惠，但是如果公司要求他们减少用电量时他们必须同意这一要求。Timko Produces 公司在星期四中午12点减少用电量。为了确定节省的用电量，公用事业公司必须估计 Timko 公司在不停电时的用电量。停电时间从中午12点到晚上8点。停电前72小时的 Timko 公司用电数据是可以得到的。

时间段	星期一	星期二	星期三	星期四
上午0~4点	—	19 281	31 209	27 330
上午4~8点	—	33 195	37 014	32 715
上午8~12点	—	99 516	119 968	152 465
下午12~4点	124 299	123 666	156 033	
下午4~8点	113 545	111 717	128 889	
下午8~12点	41 300	48 112	73 923	

a. 在24小时期间内是否存在季节影响？

b. 计算4小时期间的季节指数。

c. 利用季节指数调整的趋势来估计 Timko 公司在停电期间的正常用电量。

小结

本章我们介绍了时间序列分析和预测的基本方法。首先，我们说明了常常通过绘制时间序列图来确定时间序列的基本模式。数据模式可分为几种类型，包括水平模式、趋势模式和季节模式。我们以时间序列所呈现的这些模式为依据，讨论了预测方法。

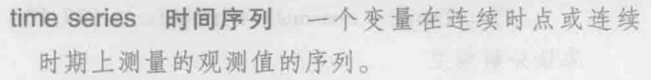
对一个具有水平模式的时间序列，我们说明了如何用移动平均法和指数平滑法进行预测。移动平均法计算过去数据值的平均数，然后用这个平均数作为下一个时期的预测值。在指数平滑法中，我们用时间序列过去数值的加权平均数来计算预测值。当水平模式移动到不同的水平，并且继续保持水平模式时，这些方法也同样适用。

在确定使用什么预测方法时的重要因素涉及的是方法的准确性。我们讨论了三种预测精度的测度：平均绝对误差（MAE）、均方误差（MSE）和平均绝对百分数误差（MAPE）。每一种测度的目的是为了确定，某种特定的预测方法在多大程度上能再现已取得的时间序列数据。通过选择对已知数据有最佳精度的方法，我们期望增加获得未来时期更好预测值的可能性。

对于只有长期线性趋势的时间序列，我们说明了如何用简单时间序列回归来进行趋势推测。我们还讨论了一种能用来预测具有线性趋势时间序列的扩展的单一指数平滑法，该方法被称为 Holt 线性指数平滑法。对于具有曲线或非线性趋势的时间序列，我们介绍了多元回归能用来拟合数据的二次趋势方程或指数趋势方程。

对于具有季节模式的时间序列，我们介绍了在多元回归模型中如何使用虚拟变量，以及多元回归模型如何用于建立有季节影响的估计的回归方程。然后，通过介绍如何将处理季节性的虚拟变量方法和处理线性趋势的时间序列回归方法组合起来，我们将回归方法扩展到同时拥有季节和线性趋势影响的时间序列。

在本章的最后一节，我们介绍了如何用时间序列分解将一个时间序列分隔或分解为季节和趋势成分，然后得到消除季节影响的时间序列。我们介绍了如何计算一个乘法模型的季节指数，如何用季节指数得到消除季节影响的时间序列，以及如何对消除季节影响的数据使用回归方程来估计趋势成分。当趋势和季节成分同时存在时，建立预测的最后一步是用季节指数调整趋势推测值。



关键术语

time series 时间序列 一个变量在连续时点或连续时期上测量的观测值的序列。

time series plot 时间序列图 时间序列图是时间和时间序列变量之间关系图形的表述。时间位于横轴，时间序列值位于纵轴。

horizontal pattern 水平模式 当数据围绕着一个不变的均值上下波动时，则存在水平模式。

stationary time series 平稳时间序列 统计性质与时间独立的时间序列。平稳时间序列过程产生的数据有一个不变的均值，而且时间序列的变异性随着时间的推移保持不变。

trend pattern 趋势模式 如果时间序列图在一段较长的时间内呈现出逐步改变或移动到相对较高的或较低的值，则存在趋势模式。

seasonal pattern 季节模式 如果时间序列图在连续的时期内而呈现重复模式，则存在季节模式。连续的时期常常指一年的周期，这也是季节模式名称的由来。

cyclical pattern 循环模式 如果时间序列图显示出持续时间超过一年的在趋势线的上下交替的点序列，则存在循环模式。

forecast error 预测误差 时间序列实际值与预测值之间的差。

mean absolute error (MAE) 平均绝对误差 预测误差绝对值的平均数。

mean squared error (MSE) 均方误差 预测误差平方和的平均数。

mean absolute percentage error (MAPE) 平均绝对百分数误差 百分数预测误差绝对值的平均数。

moving average 移动平均法 用时间序列中最近 k 期数据值的平均数作为下一个时期预测值的一种预测方法。

weighted moving average 加权移动平均法 对时间序列最近 k 期数据值选择不同的权重，然后计算加权平均数的一种预测方法。权重之和必须等于1。

exponential smoothing 指数平滑法 用过去时间序列值的加权平均数作为预测值的一种预测方法；它是加权移动平均方法的特殊情形，我们只选择一个权重——最近时期观测值的权重。

smoothing constant 平滑常数 指数平滑模型中的参数，在计算预测值时，它是最近时间序列数值的指定权重。

linear exponential smoothing 线性指数平滑 单一指数平滑法的扩展，它用两个平滑常数能对具有线性趋势的时间序列进行预测。

time series decomposition 时间序列分解 用于将一个时间序列分隔或分解出季节和趋势成分的时间序列方法。

additive decomposition model 加法分解模型 在加法模型中，时间序列 t 期的实际值是趋势成分值、季节成分值和不规则成分值之和。

multiplicative decomposition model 乘法分解模型 在乘法模型中，时间序列 t 期的实际值是趋势成分值、季节成分值和不规则成分值的乘积。

deseasonalized time series 消除季节影响的时间序列 是指剔除了季节影响的时间序列，其方法是将每一个原始时间序列观测值除以相应的季节指数。

重要公式

K 阶移动平均预测

$$F_{t+1} = \frac{\text{最近}\,k\,\text{期数据值之和}}{k} = \frac{Y_t + Y_{t-1} + \cdots + Y_{t-k+1}}{k} \tag{17-1}$$

指数平滑预测

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t \tag{17-2}$$

线性趋势方程

$$T_t = b_0 + b_1 t \tag{17-4}$$

式中

$$b_1 = \frac{\sum_{t=1}^{n}(t-\bar{t})(Y_t-\bar{Y})}{\sum_{t=1}^{n}(t-\bar{t})^2} \tag{17-5}$$

$$b_0 = \bar{Y} - b_1\bar{t} \tag{17-6}$$

Holt 线性指数平滑

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + b_{t-1}) \tag{17-7}$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)(b_{t-1}) \quad (17\text{-}8)$$

$$F_{t+k} = L_t + b_t k \quad (17\text{-}9)$$

二次趋势方程

$$T_t = b_0 + b_1 t + b_2 t^2 \quad (17\text{-}10)$$

指数趋势方程

$$T_t = b_0 (b_1)^t \quad (17\text{-}11)$$

加法分解模型

$$Y_t = \text{Trend}_t + \text{Seasonal}_t + \text{Irregular}_t \quad (17\text{-}13)$$

乘法分解模型

$$Y_t = \text{Trend}_t \times \text{Seasonal}_t \times \text{Irregular}_t \quad (17\text{-}14)$$

补充练习

42. 下表是2010～2012年9个季度中，一个投资组合中股票所占的比例。

季度	股票所占比例（%）	季度	股票所占比例（%）
2010年第一季度	29.8	2011年第一季度	32.2
2010年第二季度	31.0	2011年第二季度	31.5
2010年第三季度	29.9	2011年第三季度	32.0
2010年第四季度	30.1	2011年第四季度	31.9
		2012年第一季度	30.0

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 用指数平滑法来预测这个时间序列，分别考虑用平滑常数 $\alpha=0.2$，0.3和0.4。哪一个平滑常数提供更准确的预测？

c. 2009[㊀]年第二季度投资组合中股票所占比例的预测值是多少？

44. 为了避免计息支票账户每月的服务费，客户必须保留最低的日平均余额。2008年Barkrate在25个最大的都市地区调查了249家银行和储蓄结构，表明为了避免月服务费你必须保留的平均余额为3 462美元。由于平均服务费用为11.97美元，平均利息率仅为0.24%，如果银行的信用贷款额度基本上等于避免每月的服务费所需要的月平均余额的话，具有计息支票账户的客户没有得到太多的收益（Bankrate website，2008年10月27日）。下表是2001～2008年为了避免支付每月的服务费所需要的最低平均余额。

年份	余额（美元）	年份	余额（美元）
2001	2 435	2005	2 294
2002	2 593	2006	2 660
2003	2 258	2007	3 317
2004	2 087	2008	3 462

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 用Minitab或Excel，建立这个时间序列的线性趋势方程。计算2009年为了避免每月的服务费需要的平均余额的估计值。

c. 用Minitab或Excel，建立这个时间序列的二次趋势方程。计算2009年为了避免每月的服务费需要的平均余额的估计值。

d. 利用MSE，哪个方法对历史数据提供更准确的预测？

e. 对这些数据，你建议是用线性趋势方程还是二次曲线方程建立2009年的预测值？请解释。

46. 艾奥瓦州达文波特的Mayfair百货商店由于密西西比河洪水造成的损失，被迫在7～8月停业。该商店正试图确定由此而损失的销售额。1～6月的销售额数据如下。

月份	销售额（1 000美元）	月份	销售额（1 000美元）
1	185.72	4	210.36
2	167.84	5	255.57
3	205.11	6	261.19

a. 利用 $\alpha=0.4$ 的指数平滑法预测7月和8月的销售额（提示：用7月的预测值作为7月的实际销售额来预测8月的销售额）。使用指数平滑法，对未来超过一个时期的预测，进行评论。

b. 利用趋势推测法预测7月和8月的销售额。

c. 以7月和8月所损失的销售额240 000美元为依据，Mayfair的保险公司提出了一个理赔方案。这个数据合理吗？如果不合理，你认为合理的理赔总额应为多少？

48. Costello音乐公司开业至今已有5年。在这期间，钢琴的销售量从第一年的12架增加到最近的76架。公司的老板Fred Costello想预测来年的钢琴销售量。有关的历史资料如下。

年份	1	2	3	4	5
销售量	12	28	34	50	76

㊀ 原书此处有误，应该为2012年。——译者注

a. 绘制时间序列图，数据中呈现何种类型的模式？

b. 建立这个时间序列的线性趋势方程。该公司每年实现的销售量平均增加多少？

c. 预测第6年和第7年的销售量。

50. 参见第49题的Costello音乐公司问题。

a. 用时间序列分解方法，计算4个季度的季节指数。

b. 什么时候Costello音乐公司会经受最大的季节影响？这个季节影响合理吗？请解释。

52. Hudson Marine公司在过去7年中是C&D航海无线电设备的授权代理商。下表是每年无线电设备的销售量的数据。

年份	1	2	3	4	5	6	7
销售数量	35	50	75	90	105	110	130

a. 绘制时间序列图，是否呈现出线性趋势？

b. 用Minitab或Excel，建立这个时间序列的线性趋势方程。

c. 利用（b）中所建立的线性趋势方程，计算第8年的年销售量预测值。

54. 参见第53题的Hudson Marine问题。

a. 计算这个时间序列的中心化移动平均数。

b. 绘制能同时显示中心化移动平均数和原始时间序列的时间序列图。讨论原始时间序列图和中心化移动平均时间序列之间的差异。

c. 计算4个季度的季节指数。

d. 什么时候Hudson Marine公司会经受最大的季节影响？这个季节影响合理吗？请解释。

案例17-1 预测食品和饮料的销售

在佛罗里达州迈尔斯堡附近的卡普蒂瓦岛上的Vintage饭店，由Karen Payne拥有并经营。饭店刚刚经营了3年。作为一家专门从事海鲜的高档餐饮企业，自从Karen的饭店开业以来，她一直在为Vintage树立声誉。经过Karen及其员工的努力，她的饭店已经成为岛上最好的且营业额增长最快的饭店之一。

为了更好地规划饭店未来的发展，Karen需要建立一个系统以提前一年预测食品和饮料的每个月的销售额。表17-26是开业前3年的食品和饮料的销售总额数据。

表17-26 Vintage饭店的食品和饮料的销售总额 （单位：1 000美元）

月份	第1年	第2年	第3年
1	242	263	282
2	235	238	255
3	232	247	265
4	178	193	205
5	184	193	210
6	140	149	160
7	145	157	166
8	152	161	174
9	110	122	126
10	130	130	148
11	152	167	173
12	206	230	235

管理报告

对Vintage饭店的销售额数据进行分析，请为Karen准备一份报告，该报告总结了你的发现、预测和建议。包括：

1. 时间序列图。对时间序列的基本模式作出评论。

2. 数据的季节性分析。计算每个月的季节指数，并对季节性的高销售月和低销售月做出评论。季节指数是否有直观上的意义？请讨论。

3. 消除时间序列的季节影响。在消除季节影响的时间序列中，是否呈现出任何趋势？

4. 利用时间序列分解法，预测第4年1～12月的销售额。

5. 利用虚拟变量回归法，预测第4年1～12月的销售额。

6. 在你报告的附录中，给出你的计算汇总表和图。

假设第4年1月份的销售额为295 000美元，你的预测误差是多少？如果这个误差比较大，Karen可能会对你的预测值和实际销售额之间的差异生产疑惑，你将如何消除她对预测方法的怀疑？

案例17-2 预测损失的销售额

当8月31日飓风侵袭时，Carlson百货商店遭受了严重的损失。商店被迫关闭了4个月（9～12月），目前，Carlson正和保险公司就停业期间所损失的销售总额进行谈判。必须解决的两个关键性问题是：①如果没有遭到飓风的侵袭，Carlson将会有多少销售额；②由于商业活动的增加得到了额外的销售额，Carlson是否会有权获得任何赔偿。该县得到了超过80亿美元的联邦灾难救济金和保险金，这导致百货商店销售额的增加和其他众多的商业活动。

表17-27是Carlson遭到飓风袭击之前48个月的销售额，表17-28是该县所有百货商店遭到飓风袭击之前48个月的销售总额，同时也给出了在Carlson百货商店停业的4个月中的销售总额。Carlson的管理人员要求你分析这些数据，并且估计Carlson百货商店9～12月损失的销售额。他们同时要求你确定，在同一时期是否存在由于飓风而产生的额外销售额。如果存在的话，Carlson除了得到正常销售额的补偿外，还有权得到额外销售额是赔偿补偿。

表17-27 Carlson百货商店的销售额 （单位：10亿美元）

月份	年份1	年份2	年份3	年份4	年份5
1		1.45	2.31	2.31	2.56
2		1.80	1.89	1.99	2.28
3		2.03	2.02	2.42	2.69
4		1.99	2.23	2.45	2.48
5		2.32	2.39	2.57	2.73
6		2.20	2.14	2.42	2.37
7		2.13	2.27	2.40	2.31
8		2.43	2.21	2.50	2.23
9	1.71	1.90	1.89	2.09	
10	1.90	21.30	2.29	2.54	
11	2.74	2.56	2.83	2.97	
12	4.20	4.16	4.04	4.35	

表17-28 该县百货商店的销售额 （单位：10亿美元）

月份	年份1	年份2	年份3	年份4	年份5
1		46.80	46.80	43.80	48.00
2		48.00	48.60	45.60	51.60
3		60.00	59.40	57.60	57.60
4		57.60	58.20	53.40	58.20
5		61.80	60.60	56.40	60.00
6		58.20	55.20	52.80	57.00
7		56.40	51.00	54.00	57.60
8		63.00	58.80	60.60	61.80
9	55.80	57.60	49.80	47.40	69.00
10	56.40	53.40	54.60	54.60	75.00
11	71.40	71.40	65.40	67.80	85.20
12	117.60	114.00	102.00	100.20	121.80

管理报告

请为Carlson百货商店的管理人员准备报告，该报告总结了你的发现、预测和建议，包括：

1. 假如没有飓风袭击，估计Carlson的销售额。
2. 假如没有飓风袭击，估计全县百货商店的销售额。
3. 估计Carlson百货商店9～12月停业期间损失的销售额。

除此之外，利用全县百货商店9～12月的实际销售额和（2）的估计，对飓风产生额外销售额的提法，作出同意或反对的判断。

第18章 非参数方法

CHAPTER 18

实践中的统计

West Shell Realtors 公司[a]

俄亥俄州，辛辛那提

West Shell Realtors 公司成立于1958年，刚成立时只有一间办事处和三个营销人员。1964年公司开始实施一项长期扩张方案，每年都有新的办事处加入。几年以后，West Shell Realtors 已经是大辛辛那提地区最大的房地产经纪公司之一，而且在俄亥俄州的西南部、印第安纳州的东南部以及肯塔基州的北部都设有办事处。

统计分析可以帮助房地产公司，如 West Shell Realtors 监督其销售业绩。West Shell Realtors 公司的每个办事处及总公司每个月都编制业绩报表。有关总销售金额、已销售单元个数及每个单元的平均售价的统计汇总数据，对于办事处的经理和公司高层领导掌握公司的经营过程和难点所在至关重要。

除了对每个月的业绩进行总结之外，该公司还利用统计方法来指导公司的发展计划和战略。West Shell Realtors 正在执行一项有计划的扩张战略。每当一项扩张计划需要成立一个新的销售办事处时，办事处的选址问题就会成为公司关心的焦点。房屋的销售价格、周转率和预计的销售量都是评价和比较办事处选址时需要考虑的数据。

例如，该公司选择了两个郊区——克利夫顿和罗斯劳尔作为新的办事处首选地点。在比较两个地区时需要考虑很多因素，其中包括房屋的销售价格。公司利用非参数统计方法来帮助确定这两个地区销售模式的不同之处。

公司从克利夫顿选取了25笔销售业务，从罗斯劳尔选取了18笔销售业务组成样本，然后选择曼-惠特尼-威尔科克森秩和检验法对销售价格的差异进行统计检验。在显著性水平0.05下，曼-惠特尼-威尔科克森检验并没有拒绝这两个地区销售价格相等这一原假设。于是，公司在地点选择过程中可以集中考察房屋销售价格以外的选择标准。

在本章中，我们将学习如何应用诸如曼-惠特尼-威尔科克森检验之类的非参数统计检验，同时，我们还将讨论有关这些检验的正确解释。

到目前为止，本书统计推断中所介绍的方法都是通常所知的**参数方法**（parametric method）。这些方法通常始于假定总体的概率分布服从正态分布，基于这个假定，统计学家可以得到用于推断一个或多个总体参数（例如总体均值μ、总体标准差σ）的抽样分布。例如，在第9章，我们介绍了假定总体服从正态概率分布，其参数μ和σ未知情况下对总体均值进行统计推断的方法。用样本标准差s来估计总体标准差σ，对总体均值进行推断的统计量服从t分布。因此，对正态总体均值可以用t分布确定置信区间和进行检验。

在本章中，我们将介绍用于对总体进行推断的**非参数方法**（nonparametric method），非参数方法对总体概率分布形式的假定没有要求，因此，非参数方法也被称为**无分布方法**（distribution-free method）。

绝大多数的参数统计方法要求数量型数据，而非参数方法则允许基于分类型或品质型数据进行推断。用于非参数方法的计算一般与分类型数据有关。即使数据是数值型的，为进行非参数检验，我们也要将其转化为分类型数据。本章的第18.1节我们将演示如何用二项分布对两类数据以及总体中位数进行推断；在接下来的三节中，我们演示如何用秩-顺序数据对两个或多个总体进行非参数检验。最后一节，我们用秩-顺序数据对两个变量计算等级相关系数。

18.1 符号检验

符号检验（sign test）是假设检验中一种多用途的非参数方法，其利用$p=0.50$的二项分布作为抽样分布。它对总体的分布没有假定的要求。本节我们将介绍符号检验的两个应用：一个是关于总体中位数的检验；另一个是两个匹配总体之差的检验。

[a] 作者感谢为“实践中的统计”提供了本案例的 West Shell Realtors 公司的 Rodney Fightmaster 先生。

18.1.1 总体中位数假设检验

在第 9 章中，我们阐述了如何利用假设检验来对总体均值进行检验。本节我们将阐述如何用符号检验来对总体中位数进行假设检验。如果我们考虑一个总体，其中没有任何一个数值恰好等于中位数，则中位数是中心趋势的度量，它将总体化分成为两部分，其中 50% 的值大于中位数，另外 50% 的值小于中位数。每当总体是偏斜分布时，作为总体中心位置的最佳度量，中位数也常常优于平均数。符号检验提供了检验总体中位数假设的非参数方法。

表 18-1　Cape May 薯片在 10 个 Lawler 食品店的周销售额

商店编号	周销售额（美元）
56	485
19	562
36	415
128	860
12	426
63	474
39	662
84	380
102	515
44	721

为了阐明符号检验，我们考虑 Lawler 食品连锁店中 Cape May 薯片周销售额。根据制造商估计每个食品店每周销售额的中位数应为 450 美元，Lawler 的经理作出销售新薯片产品的决定。在销售该产品 3 个月后，Lawler 的经理要求进行周销售额总体中位数的假设检验。

$$H_0: \text{总位数} = 450$$
$$H_a: \text{中位数} \neq 450$$

表 18-1 是随机抽取的 10 个 Lawler 食品店一周的销售额数据。在符号检验中，我们将每一个样本观测值与总体中位数的假设值进行比较。如果观测值大于假设值，我们用加号表示；如果观测值小于假设值，我们用减号表示；如果观测值恰好等于假设值[㊀]，则我们应该将它从样本中剔除，并且在一个较小的样本容量中进行分析，只使用被标记为加号或减号的观测值。它将样本数据转化为加号或减号，因此，这种非参数方法被命名为符号检验。

考虑表 18-1 中的样本数据。第一个观测值 485 大于假设的中位数 450，记为加号；第二个观测值 562 大于假设的中位数 450，记为加号，以此类推。表 18-2 是 10 个样本观测值对应的加号或减号数据，注意有 7 个加号和 3 个减号。

表 18-2　周销售额总体中位数符号检验的 Lawler 样本数据

商店编号	周销售额（美元）	符号	商店编号	周销售额（美元）	符号
56	485	+	63	474	+
19	562	+	39	662	+
36	415	−	84	380	−
128	860	+	102	515	+
12	426	−	44	721	+

加号和减号的分配适合应用二项分布，样本容量 $n=10$ 是试验的次数，每次试验有两个可能的结果：加号或减号，而且试验是独立的。令 p 表示加号的概率，如果总体中位数为 450，则 $p=0.50$，即总体中应该有 50% 的加号和 50% 的减号。因此，就二项概率 p 而言，总体中位数的符号检验的假设

$$H_0: \text{总位数} = 450$$
$$H_a: \text{中位数} \neq 450$$

转化为下面关于二项概率 p 的假设

$$H_0: p = 0.50$$
$$H_a: p \neq 0.50$$

如果 H_0 没有被拒绝，我们不能得出 $p \neq 0.5$ 的结论，进而我们不能得出总体中位数 ≠450 的结论。然而，如果 H_0 被拒绝，我们能得出 $p \neq 0.5$，进而我们能得出总体中位数 ≠450。

由于 $n=10$ 个食品店或 $n=10$ 次试验，且 $p=0.50$，我们用附录 B 中的表 5 得到在假设 H_0 为真的条件下加号个数的二项分布的概率值，这些概率值如表 18-3 所示。[㊁]图 18-1 是这个二项概率分布的图形表示。

㊀ 观测值等于假设值将被剔除，只对具有加号或减号的观测值进行分析。

㊁ 当样本容量小于或等于 20 时，在附录 B 的表 5 中可以查到具体的二项概率。用 Excel 或 Minitab 可以得到任何样本容量的二项概率。

表 18-3 $n=10$，$p=0.50$ 的二项概率

加号个数	概率
0	0.001 0
1	0.009 8
2	0.043 9
3	0.117 2
4	0.205 1
5	0.246 1
6	0.205 1
7	0.117 2
8	0.043 9
9	0.009 8
10	0.001 0

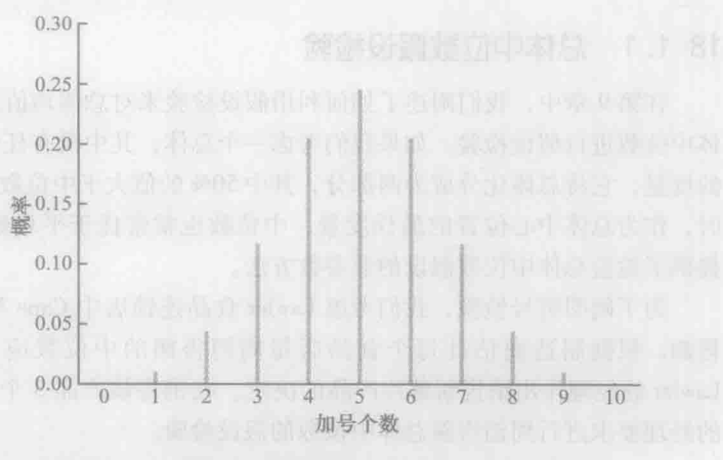


图 18-1 $n=10$、$p=0.50$ 的加号个数的二项抽样分布

下面我们演示如何用二项分布进行总体中位数的假设检验。我们使用 0.10 为检验的显著性水平。由于样本数据中观测到的加号个数为 7，位于二项分布的上侧，我们计算加号大于或等于 7 的概率，即加号个数为 7，8，9 或 10 的概率。将表 18-3 中的这些概率相加，我们得到 $0.1172+0.0439+0.0098+0.0010=0.1719$。由于是双侧假设检验，将这个上侧概率加倍，得到 p-值为 $2\times0.1719=0.3438$。由于 p-值 $>\alpha$，所以我们不能拒绝假设 H_0。对于二项概率，我们不能拒绝 H_0：$p=0.50$，因此我们不能拒绝总体中位数为 450 的假设。

在本例中，总体中位数的假设检验是一个双侧检验，但是，关于总体中位数的符号检验也有可能是单侧。例如我们所用的假设是上侧检验，因此，原假设和备择假设可表示为

$$H_0: 总位数 \leqslant 450$$
$$H_a: 中位数 > 450$$

相应的 p-值等于样本中加号个数大于或等于 7 的二项概率。这个单侧 p-值等于 $0.1172+0.0439+0.0098+0.0010=0.1719$。如果转化为下侧检验的例子，$p$-值将是加号个数为小于或等于 7 的概率。

我们刚才介绍了使用 $p=0.50$ 的二项分布的应用。当样本容量等于或小于 20 时，用附录 B 表 B-5 给出的二项概率能计算 p-值。对于较大的样本容量，我们根据二项分布的正态分布近似来计算 p-值，这使得计算变得迅速且容易。下面的例子阐明符号检验的大样本应用。

一年前新建住宅价格的中位数为 236 000 美元。但是，当前经济低迷，使得房地产公司用当前的住宅销售的样本数据来确定今年新建住宅价格的总体中位数是否比一年前有所下降。

新建住宅价格的总体中位数的假设检验如下：

$$H_0: 中位数 \geqslant 236\,000\ 美元$$
$$H_a: 中位数 < 236\,000\ 美元$$

我们在显著性水平 0.05 下进行这一检验。

由 61 所最近新建住宅组成的一个样本，22 所住宅的价格高于 236 000 美元，38 所住宅的价格低于 236 000 美元，还有 1 所住宅的价格恰好等于 236 000 美元。在剔除价格等于假设中位数价格 236 000 美元的那所住宅后。符号检验中有 22 个加号、38 个减号，样本容量为 60。

总体中位数大于或等于 236 000 美元的原假设表示成二项分布的假设为 H_0：$p\geqslant0.5$。如果 H_0 为真，我们期望加号个数为 $0.50\times60=30$（所）住宅。样本结果表明有 22 个加号，位于二项分布的下侧，则 p-值是当 $p=0.50$ 时，加号个数小于或等于 22 的概率。虽然计算 0，1，2，…，22 的二项概率的精确值并求和是可能的，但我们用二项分布的正态分布近似使计算变得很容易。对这个近似，正态分布的均值和标准差如下所示：

当 H_0：$p=0.50$ 时，加号个数的抽样分布的正态近似

$$\text{均值}: \mu = 0.50n \tag{18-1}$$

$$\text{标准差}: \sigma = \sqrt{0.25n} \tag{18-2}$$

分布形式：当 $n>20$ 时，近似正态分布。

对于 $n=60$ 和 $p=0.50$，利用式（18-1）和式（18-2），加号个数的抽样分布能近似为一个正态分布，其中

$$\mu = 0.50n = 0.50 \times 60 = 30$$

$$\sigma = \sqrt{0.25n} = \sqrt{0.25 \times 60} = 3.873$$

现在我们用正态分布近似加号个数小于或等于22的二项概率。在我们进行近似之前，需要牢记二项分布是离散型的，而正态分布是连续型的，因此，解决这一问题，用区间（21.5，22.5）上的正态概率来计算个数为22的二项概率。0.5被称为连续性校正因子。于是，计算加号个数小于或等于22的 p-值，近似为服从均值 $\mu=30$，标准差 $\sigma=3.873$ 的正态随机变量 x 小于或等于22.5的概率，图18-2是这个 p-值的图示。

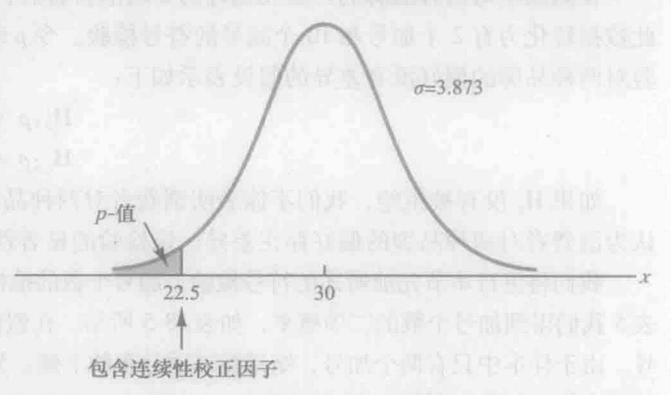


图18-2 新建住宅价格中位数的符号检验中 p-值的正态分布近似

利用正态分布，我们得到 p-值如下：

$$p\text{-值} = P(x \leqslant 22.5) = P\left(z \leqslant \frac{22.5-30}{3.873}\right) = P(z \leqslant -1.94)$$

利用正态概率分布表，我们看到 $z=-1.94$ 的累积概率提供的 p-值 $=0.0262$。由于 $0.0262<0.05$，我们拒绝原假设，并且得出结论：新建住宅价格的中位数低于一年前的中位数价格236 000美元。

注释和评论

1. 用来说明总体中位数的假设检验的例子涉及周销售额和住宅价格数据，这些变量的概率分布类型通常是非对称的，而且大多数常常是右偏的。在这种情形下，总体中位数而非总体均值成为中心位置的首选测度。一般地，当总体非对称时，总体中位数的非参数符号检验通常是较合适的统计检验。
2. 符号检验的二项抽样分布可以用来计算总体中位数的置信区间估计。然而，这个计算相当复杂而且极少能用手工完成。利用如Minitab等统计软件包能够得到总体中位数的置信区间，在附录18A中描述了确定总体中位数置信区间的Minitab步骤。用本节的住宅价格的例子，Minitab给出了新建住宅价格中位数的置信区间为（183 000美元，231 000美元）。

18.1.2 匹配样本的假设检验

在第10章，我们介绍了匹配样本实验设计，n 个实验单位中的每一个提供一对观测值，其中一个来自总体1，另一个来自总体2。使用数量型数据并且假设匹配观测值之差服从正态分布，则可以使用 t 分布对两个总体均值之差进行推断。

在接下来的例子中我们用非参数符号检验来分析匹配样本数据。与 t 分布方法不同（其要求数量型数据，且假定二者之差服从正态分布），符号检验既能分析分类型数据，也能分析数量型数据，并且对二者之差的分布没有假定。在市场调查中要求 n 个潜在的消费者比较两个品牌的产品（如咖啡、软饮料或洗涤剂）是匹配样本的典型设计。不需要得到每个消费者对品牌偏好的数值测度，只要求每个消费者陈述对品牌偏好。考虑下面的例子。

阳光海岸农场生产一种名为“柑橘谷”的橙汁。柑橘谷的主要竞争对手来自名为“热带橘”的橙汁饮品。在消费者对这两种品牌偏好的比较中，14名消费者面对的两种橙汁产品都没有标出品牌，每一名消费者第一次品尝

的品牌都是随机选择的。如果消费者更偏好柑橘谷，则用加号来表示；如果消费者更偏好热带橘，则用减号来表示；如果消费者对两种产品的偏好无差异，则没有符号表示。表 18-4 是参与研究的 14 名消费者的数据。

表 18-4　阳光海岸农场口味检测的偏好数据

消费者	偏好的品牌	符号	消费者	偏好的品牌	符号
1	热带橘	−	8	热带橘	−
2	热带橘	−	9	热带橘	−
3	柑橘谷	+	10	无偏好	
4	热带橘	−	11	热带橘	−
5	热带橘	−	12	柑橘谷	+
6	无偏好		13	热带橘	−
7	热带橘	−	14	热带橘	−

在剔除了对两种品牌的产品无偏好的 2 名消费者后，$n=12$ 名消费者都能表达出对两种品牌之一的偏好，因此数据转化为有 2 个加号和 10 个减号的符号检验。令 p 表示消费者总体中偏好柑橘谷的所占的比例，我们想要检验对两种品牌的偏好没有差异的假设表示如下：

$$H_0: p = 0.50$$
$$H_a: p \neq 0.50$$

如果 H_0 没有被拒绝，我们不能表明消费者对两种品牌的偏好存在差异。然而，如果 H_0 被拒绝，我们就可以认为消费者对两种品牌的偏好存在差异。该检验的显著性水平为 0.05。

我们将进行本节先前所述的符号检验。加号个数的抽样分布服从 $p=0.50$ 和 $n=12$ 的二项分布。利用附录 B 中的表 5 我们得到加号个数的二项概率，如表 18-5 所示。在假设 H_0 为真的条件下，我们期望有 $0.50n=0.50\times12=6$ 个加号。由于样本中只有两个加号，结果在二项分布的下侧。为计算双侧检验的 p-值，我们先计算加号个数小于或等于 2 的概率然后加倍。利用表 18-5 中的 0，1 和 2 的二项概率，p-值 $=2\times(0.0002+0.0029+0.0161)=0.0384$，由于 $0.0384<0.05$，所以我们拒绝假设 H_0，口味检测提供证据表明消费者对两种品牌的橙汁的偏好存在差异。我们有必要向阳光海岸农场提出建议，因为竞争对手热带橘产品更受青睐，因此阳光海岸农场应该寻求处理此问题的战略。

表 18-5　$n=12$，$p=0.50$ 的二项概率

加号个数	概率	加号个数	概率	加号个数	概率
0	0.000 2	5	0.193 4	10	0.016 1
1	0.002 9	6	0.225 6	11	0.002 9
2	0.016 1	7	0.193 4	12	0.000 2
3	0.053 7	8	0.120 8		
4	0.120 8	9	0.053 7		

与符号检验的其他使用情形类似，根据实际应用也可能使用单侧检验。另外，当样本容量变大时，正如本节先前所述，二项分布的正态分布近似可使计算变得很容易。在阳光海岸农场匹配样本的符号检验中使用的是分类型的偏好数据，匹配样本的符号检验也可以使用数量型数据。如果匹配数据二者之差不服从正态分布且是偏态分布时，匹配样本的符号检验尤其有意义，此时，正的差值记为加号，负的差值记为减号，差值为 0 将从样本中剔除。符号检验的计算过程同前所述。

练习

方法

☆2. 10 个人参加了两种品牌的产品的口味测试。样本结果显示，7 个人偏好品牌 A，2 个人偏好品牌 B，1 个人没有陈述偏好。在显著性水平 $\alpha=0.05$ 下，对两个品牌偏好的差异进行显著性检验。你的结论如何？

应用

4. 最大的 50 只股票型共同基金的资产净值的中位数为 150 亿美元（《华尔街日报》，2009 年 3 月 2 日）。由最大的 50 只债券型共同基金中的 10 只组成一个样本，其数据如下：

债券型基金	资产净值
Fidelity Intl Bond	61
Franklin CA TF	117
American Funds	224
Vanguard Short Term	96
PIMCO：Real Return	49
T Rowe Price New Income	69
Vanguard GNMA	150
Oppenheimer Intl Bond	66
Dodge & Cox Income	145
iShares：TIPS Bond	96

在显著性水平 $\alpha=0.05$ 下，利用中位数能否得出结论：债券型共同基金的资产净值小于股票型共同基金？

a. 检验的假设是什么？

b. p-值是多少？你的结论如何？

6. 美国家庭年收入的中位数是 56 200 美元（*The New York Times Almanac*，2008 年）。下面是居住在伊利诺伊州芝加哥的 50 个家庭组成的样本数据（单位：1 000 美元）。在显著性水平 $\alpha=0.05$ 下，利用样本数据能否得出结论：居住在芝加哥的家庭年收入的中位数大于 56 200 美元。你的结论如何？

66.3	60.2	49.9	75.4	73.7
65.7	61.1	123.8	57.3	48.5
74.0	146.3	92.2	43.7	86.9
59.7	64.2	56.2	48.9	109.6
39.8	60.9	79.7	42.3	52.6
60.9	43.5	61.7	54.7	95.2
70.4	43.8	57.8	83.5	56.5
51.3	42.9	87.5	43.6	67.2
48.7	79.1	61.9	53.4	56.2
57.0	49.6	109.5	42.1	74.6

8. Pew 研究中心进行的一项调查是询问成年人，他们理想的生活地区是节奏较快还是较慢的地区（《今日美国》，2009 年 2 月 13 日）。最初的 16 名回答者的样本中，有 4 人喜欢生活节奏较快的地区，11 人喜欢生活节奏较慢的地区，1 人无所谓。

a. 这些数据能否得出结论：人们对生活节奏较快和较慢的偏好之间存在差异？在显著性水平 $\alpha=0.05$ 下，你的结论如何？

b. 考虑 16 人的样本，喜欢生活节奏较快的比例是多少？喜欢生活节奏较慢的比例是多少？你对该研究有何建议？

10. 尼尔森媒体研究表明，*American Idol* 和 *Dancing with the Stars* 是黄金时间收视率最高的两个电视节目（《今日美国》，2008 年 4 月 14 日）。假设在一项地方电视台偏好的调查中，要求750 个被调查者说出他们喜欢的黄金时间电视节目，其中 330 个人选择了 *American Idol*，270 个人选择了 *Dancing with the Stars*，还有150 个人选择了其他的电视节目。在显著性水平 0.05 下检验假设：对 *American Idol* 和 *Dancing with the Stars* 两个电视节目的偏好不存在差异。你的结论如何？

18.2 威尔科克森符号秩检验

在第 10 章中，我们介绍了匹配样本实验设计，n 个实验单位中的每一个提供一对观测值，其中一个来自总体 1，另一个来自总体 2。该实验的参数检验要求数量型数据，而且假定配对观测值之差服从正态分布，则可以利用 t 分布对两个总体均值之差进行推断。

威尔科克森符号秩检验（Wilcoxon signed-rank test）是分析匹配样本实验数据的非参数方法。检验使用数量型数据，但不要求假定配对观测值之差服从正态分布，只需要假定配对观测值之差具有对称的分布㊀。即使两个总体的形态相同，关注点是确定两个总体的中位数是否有差异。用下面的例子来说明威尔科克森符号秩检验。

考虑某个制造企业正在尝试确定两种生产方法在完成任务时间上是否存在差异。使用匹配样本设计，随机选择 11 个工人的 2 次完成任务时间，1 次使用方法 A，1 次使用方法 B。工人首先使用的生产方法是随机挑选的。两种方法的完成任务时间以及完成时间的差异数据如表 18-6 所示。正的差异表明方法 A 需要更多的时间，而负的差异则表明方法 B 需要更多的时间。这些数据是否表明两种方法在完成任务时间上存在显著差异呢？如果我们假设数据的差异具有对称分布，但不需要正态分布，可以应用威尔科克森符号秩检验。

㊀ 如果二者之差的总体分布是偏斜分布，则推荐使用第 18.1 节介绍的匹配样本的符号检验。

表 18-6 完成生产任务的时间（分钟）

工人	方法 A	方法 B	差	工人	方法 A	方法 B	差
1	10.2	9.5	0.7	7	10.6	10.5	0.1
2	9.6	9.8	-0.2	8	10.0	10.0	0.0
3	9.2	8.8	0.4	9	11.2	10.6	0.6
4	10.6	10.1	0.5	10	10.7	10.2	0.5
5	9.9	10.3	-0.4	11	10.6	9.9	0.8
6	10.2	9.3	0.9				

特别地，我们用威尔科克森符号秩检验对两种方法完成时间的中位数之差进行检验[㊀]，假设如下：

$$H_0: 方法\ A\ 的中位数 - 方法\ B\ 的中位数 = 0$$

$$H_a: 方法\ A\ 的中位数 - 方法\ B\ 的中位数 \neq 0$$

如果无法拒绝 H_0，我们将不能得出两种方法的完成任务时间的中位数存在差异。但是，如果 H_0 被拒绝，则我们可以得出两种方法在完成任务时间的中位数上存在差异。我们在显著性水平 0.05 下进行检验。

威尔科克森符号秩检验的第一步剔除差异为 0[㊁] 的工人 8，然后计算剩下 10 个工人差的绝对值，如表 18-7 中的第 3 列所示。接下来将差的绝对值由低到高进行排序，如表 18-7 中的第 4 列所示。工人 7 的最小的绝对差 0.1 的秩被定为 1，工人 2 的第二小的绝对差 0.2 的秩被定为 2，这种对绝对差的排秩一直将继续下去，直到工人 6 的最大绝对差 0.9 的秩被赋为 10 为止。工人 3 和工人 5 有相同的绝对差 0.4，0.4 的秩被定为平均秩[㊂]3.5；与此同时，工人 4 和工人 10 有相同绝对差 0.5，0.5 的秩被定为平均秩 5.5。

一旦确定了绝对差的秩，这些秩将被赋予工人原始差的符号，负的符号秩列在第 5 列，正的符号秩列在第 6 列（见表 18-7）。例如，工人 1 的差是 0.7（见第 2 列），其绝对差的秩为 8（见第 4 列），因此，工人 1 的秩在第 6 列中是正的符号秩 8。工人 2 的差为 -0.2，其绝对差的秩为 2，因此，工人 2 的秩在第 5 列是负的符号秩为 -2。继续这个过程直到产生如表 18-7 所示的负的和正的符号秩为止。

表 18-7 完成生产任务时间的绝对差的秩和符号秩

工人	差	差的绝对值	秩	符号秩 负	符号秩 正
1	0.7	0.7	8		8
2	-0.2	0.2	2	-2	
3	0.4	0.4	3.5		3.5
4	0.5	0.5	5.5		5.5
5	-0.4	0.4	3.5	-3.5	
6	0.9	0.9	10		10
7	0.1	0.1	1		1
8	0.0	0.0			
9	0.6	0.6	7		7
10	0.5	0.5	5.5		5.5
11	0.8	0.8	9		9
				正的符号秩之和	$T^+=49.5$

令 T^+ 表示正的符号秩之和，这里 $T^+=49.5$。为了进行威尔科克森符号秩检验，我们将用 T^+ 作为检验的统计量。如果两个总体中位数相等，并且匹配数据对的个数为大于或等于 10 时，T^+ 的抽样分布近似于如下的正态分布。

㊀ 本节的例子所用的观点是两总体的形态相同，如果它们仅仅是位置上的不同，就总体中位数而言，也能陈述威尔科克森符号秩检验的假设。

㊁ 剔除差为 0 的情况，分析继续在差不为 0 的容量更小的样本中进行。

㊂ 相同的绝对差被定为它们秩的平均值。

威尔科克森符号秩检验中 T^+ 的抽样分布

$$均值:\mu_{T^+}=\frac{n(n+1)}{4} \quad (18\text{-}3)$$

$$标准差:\sigma_{T^+}=\sqrt{\frac{n(n+1)(2n+1)}{24}} \quad (18\text{-}4)$$

分布形式：当 $n\geqslant10$ 时，近似正态分布。

在剔除了差为0的观测值（工人8）后，我们得到 $n=10$ 匹配数据。因此，利用式（18-3）和（18-4），我们有

$$\mu_{T^+}=\frac{n(n+1)}{4}=\frac{10\times11}{4}=27.5$$

$$\sigma_{T^+}=\sqrt{\frac{n(n+1)(2n+1)}{24}}=\sqrt{\frac{10(10+1)(20+1)}{24}}=\sqrt{\frac{2\,310}{24}}=9.810\,7$$

图 18-3 是检验统计量 T^+ 的抽样分布。

对于两种方法完成任务时间的中位数相等的假设，让我们计算其双侧 p-值。由于检验统计量 $T^+=49.5$ 在抽样分布的上侧，则我们计算上侧概率 $P(T^+\geqslant49.5)$。由于正的符号秩之和 T^+ 是离散型的，而正态分布是连续型，我们用包含连续性校正因子来得到最好的近似，因此，$T^+=49.5$ 的离散概率近似为区间（49，50）上的正态概率。$T^+=49.5$ 的概率近似为

$$P(T^+\geqslant49.5)=P\left(z\geqslant\frac{49-27.5}{9.810\,7}\right)=P(z\geqslant2.19)$$

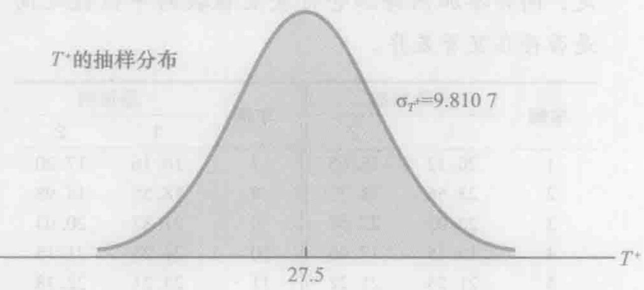


图 18-3　完成任务时间例子中的 T^+ 的抽样分布

利用标准正态分布表和 $z=2.19$，我们看到双侧 p-值为 $2\times(1-0.985\,7)=0.028\,6$。由于 p-值 $\leqslant0.05$，所以我们拒绝 H_0，故得两种方法在完成任务时间的中位数不相等的结论。由于 T^+ 在抽样分布的上侧，我们得到方法 A 导致较长的完成时间。我们期待管理部门认定，方法 B 是较快或较好的生产方法。

单侧威尔科克森符号秩检验也是可能的。例如，如果最初我们找到统计证据证实方法 A 完成任务时间的中位数较长，方法 B 完成任务时间的中位数较短，我们可以用如下的上侧假设检验：

H_0:方法 A 的中位数 − 方法 B 的中位数 $\leqslant 0$

H_a:方法 A 的中位数 − 方法 B 的中位数 > 0

拒绝 H_0 将得出方法 A 完成任务时间的中位数较长，方法 B 完成任务时间的中位数较短。下侧假设检验也是可能的。

最后一点，在第 18.1 节中我们演示了符号检验如何用于一个总体中位数和匹配样本的假设检验，本节我们说明了对匹配样本使用威尔科克森符号秩检验。然而，威尔科克森符号秩检验也可以用于一个总体中位数[⊖]的非参数检验，该检验除了要求总体分布对称之外对分布无其他假定。如果假定合适，威尔科克森符号秩检验是总体中位数的首选非参数检验。但是，如果总体是偏斜分布，则首选第 18.1 节介绍的符号检验。对于威尔科克森符号秩检验，总体中位数的观测值和假设值之间的差用于取代配对观测值之间的差，除此之外，计算过程如本节所述。练习第 17 题要求用威尔科克森符号秩检验对一个对称总体的中位数进行假设检验。

注释和评论

1. 一个总体中位数的威尔科克森符号秩检验基于假定总体分布对称。由于这个假定，总体中位数等于总体均值，因此，威尔科克森符号秩检验也可以用于一个对称总体均值的检验。

[⊖] 威尔科克森符号秩检验可以用于检验对称总体中位数的假设，如果总体是偏斜分布，则首选第 18.1 节介绍的符号检验。

2. 威尔科克森符号秩检验方法可以用来计算对称总体中位数的置信区间。然而，这个计算相当复杂而且极少用手工完成。利用如 Minitab 等统计软件包能够得到这个置信区间。

练习

应用

在下面涉及配对之差的练习中，我们认为：比较的总体有近似相同的形态，并且配对之差的分布近似对称的假设合理。

☆12. 对燃料添加剂进行检验，以确定它们对轿车每加仑行驶里程数的影响。对 12 辆车的每辆车都使用两种燃料添加剂进行检验，检测结果如下。在显著性水平 $\alpha=0.05$ 下，用威尔科克森符号秩检验来确定，两种添加剂每加仑行驶里程数的中位数之间是否存在显著差异。

车辆	添加剂		车辆	添加剂	
	1	2		1	2
1	20.12	18.05	7	16.16	17.20
2	23.56	21.77	8	18.55	14.98
3	22.03	22.57	9	21.87	20.03
4	19.15	17.06	10	24.23	21.15
5	21.23	21.22	11	23.21	22.78
6	24.77	23.80	12	25.02	23.70

14. 随机抽取了 11 个机场，搜集了 2006 年和 2007 年的航班正点率，数据如下表所示（Research and Innovative Technology Administration website，2008 年 8 月 29 日）。在显著性水平 $\alpha=0.05$ 下，检验两年航班正点率的中位数是否存在差异。你的结论如何？

机场	正点率	
	2006	2007
Boston Logan	71.78	69.69
Chicago O'Hare	68.23	65.88
Chicago Midway	77.98	78.40
Denver	78.71	75.78
Fort Lauderdale	77.59	73.45
Houston	77.67	78.68
Los Angeles	76.67	76.38
Miami	76.29	70.98
New York（JFK）	60.39	62.84
Orange	79.91	76.49
Washington（Dulles）	75.55	72.42

（续）

16. 2008 年 8 月 11 ~ 17 日，在北卡罗来纳州格林斯博罗的 Sedgefield 乡村俱乐部举行了美国职业高尔夫协会的球员锦标赛。下面是随机选取的 11 位高尔夫球员第一轮和第二轮的杆数。在显著性水平 $\alpha=0.05$ 下，确定高尔夫球员在锦标赛中第一轮和第二轮杆数的中位数是否存在显著差异，你的结论如何？

高尔夫球员	第一轮	第二轮
Marvin Laird	63	74
Jimmy Walker	70	73
Kevin Chappell	72	70
Kevin Duke	65	71
Andrew Buckle	70	74
Paul Claxton	69	73
Larry Mize	72	71
Chris Riley	68	70
Bubba Watson	70	68
Carlos Franco	71	71
Richard Johnson	72	69

18.3 曼 - 惠特尼 - 威尔科克森检验

在第 10 章中，我们介绍了利用两个独立样本进行两个总体均值之差的假设检验方法，其中一个样本来自总体 1，另一个来自总体 2。参数检验要求数量型数据，且假定总体都服从正态分布。当总体标准差 σ_1 和 σ_2 未知时，用样本标准差 s_1 和 s_2 作为 σ_1 和 σ_2 的估计量，并且利用 t 分布对两个总体均值之差进行推断。

在本节我们介绍一种基于两个独立样本的两总体之间差异的非参数方法。非参数方法的优点是它既可以使用顺序型数据[㊀]也可以使用数量型数据，而且不需要假定总体服从正态分布。检验的形式由曼、惠特尼和威尔科克森联合提出，因而它被称为曼 - 惠特尼检验和威尔科克森秩和检验。这些检验是等价的，并且提供同样的结论。在本节我们将这个非参数检验称为**曼 - 惠特尼 - 威尔科克森检验**（Mann-Whitney-Wilcoxon test，MWW）。

㊀ 顺序型数据是可以被排序的分类型数据，顺序尺度已在第 1 章的第 1.2 节中详细讨论。

我们以陈述原假设和备择假设的最一般形式开始 MWW 检验，假设如下：

H_0：两个总体相同

H_a：两个总体不相同

两个总体不相同的备择假设需要做一些说明。如何 H_0 被拒绝，我们用检验得出结论：总体不相同，并且总体 1 与总体 2 相比，倾向于提供或者较小或者较大的数值。图 18-4 是总体 1 比总体 2 提供较小数值的情形。注意并非所有来自总体 1 的数值都比来自总体 2 的小。然而，图形准确地表明 H_a 是正确的，即两个总体不相同，而且总体 1 比总体 2 提供较小的数值。在双侧检验中，我们考虑备择假设为提供或者较小或者较大数值的总体；在单侧检验可能提出的备择假设是，总体 1 与总体 2 相比，提供较小或者较大数值。

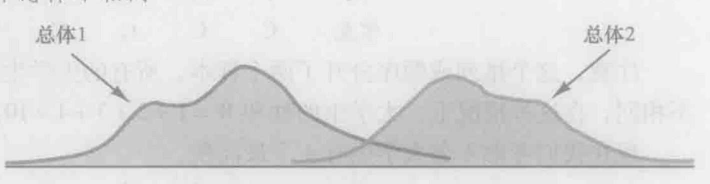


图 18-4　两个总体不相同且总体 1 倾向于提供较小数值的情形

首先我们通过对小样本秩 - 顺序数据来说明 MWW 检验。这里将使你了解如何计算秩和统计量，以及该统计量如何用于确定是否拒绝两个总体相等的原假设。在本节的稍后部分，我们将介绍基于正态分布的大样本近似，它将简化 MWW 检验所需的计算量。

我们考虑有 20 个电影银幕放映厅的 Showtime 电影院员工的工作表现。在回顾员工的表现后，影院经理在影院的年终报告中给所有 35 名员工从最好（等级 1）到最差（等级 35）排序。了解到兼职员工主要是大学生和高中生，地区经理想了解在工作表现方面大学生与高中生相比是否存在显著差异。对于在影院工作的大学生总体和高中生总体，假设陈述如下：

H_0：大学生总体和高中生总体的工作表现相同

H_a：大学生总体和高中生总体的工作表现不同

我们在显著性水平 0.05 下进行检验。

我们在 Showtime 电影院工作的学生中挑选了 4 名大学生组成一个随机样本，而另一个随机样本由 5 名高中生组成。影院经理给所有 35 名员工的综合表现排序中，也有样本中每一名员工的记录，如表 18-8 所示。选取的第一位大学生在经理的年终表现报告中排在第 15 位，选取的第二位大学生在经理的年终表现报告中排在第 3 位，等等。

表 18-8　在 Showtime 电影院工作的大学生样本和高中生样本的表现排序

大学生	经理的排名	高中生	经理的排名
1	15	1	18
2	3	2	20
3	23	3	32
4	8	4	9
		5	25

MWW 过程的下一步是对混合的样本数据由低到高进行排序。由于总共有 9 名学生，我们将表 18-8 中的表现排名数据[⊖]从 1 到 9 排序，大学生 2 的最小值 3 得到的秩为 1，大学生 4 的第二小值 8 得到的秩为 2，高中生 3 的最大值 32 得到的秩为 9。9 个学生的混合样本排秩结果列在表 18-9 中。

表 18-9　在 Showtime 电影院的 9 名学生混合样本的排秩

大学生	经理的排名	秩	高中生	经理的排名	秩
1	15	4	1	18	5
2	3	1	2	20	6
3	23	7	3	32	9
4	8	2	4	9	3
		秩和 14	5	25	8
					秩和 31

接下来是分别求出每一个样本的秩和，如表 18-9 所示。MWW 过程中可以使用任何一个样本的秩和。然而，在我们的 MWW 检验的应用中，我们遵循实践中的常用方式使用第一个样本，即 4 名大学生样本。MWW 检验的第一个样本的秩和用统计量 W 表示，因此，由表 18-9，我们得到 $W = 4 + 1 + 7 + 2 = 14$。

⊖　在本例中的数据说明 MWW 检验如何用于顺序（秩 - 顺序）数据，练习第 25 题给出了利用这类数据的另一个应用。

让我们来考虑，为什么秩和能帮助我们在两个假设，即 H_0：两个总体相同，H_a：两个总体不相同之间作出选择。令 C 表示大学生，H 表示高中生，假设 9 名学生的秩有下面顺序：4 名大学生有 4 个最低秩。

秩	1	2	3	4	5	6	7	8	9
学生	C	C	C	C	H	H	H	H	H

注意，这个排列或顺序分开了两个样本，所有的大学生都比高中生有较低的秩。有很强的迹象表明两个总体不相同，在这种情况下，大学生的秩和 $W=1+2+3+4=10$。

现在我们考虑 4 名大学生有 4 个最高秩。

秩	1	2	3	4	5	6	7	8	9
学生	H	H	H	H	H	C	C	C	C

注意，这个排列或顺序也分开了两个样本，但是此时所有的大学生都比高中生有较高的秩。它也有很强的迹象表明两个总体不相同，在这种情况下，大学生的秩和 $W=6+7+8+9=30$。于是我们看到，4 名大学生的秩和必定介于 10 ~ 30。W 的值接近于 10 就意味着大学生比高中生得到较低的秩；而 W 的值接近 30 则意味着大学生比高中生得到较高的秩。这两个极端值表明两个总体不相同。然而，如果两个总体相同，我们期望 C 和 H 的顺序交错，因此，秩和 W 将接近于两个极端值的平均值，即接近（10 + 30）/2 = 20。

假定两个总体相同，我们用计算机程序计算 9 名学生的所有可能的顺序。对每一个顺序，我们计算大学生的秩和。图 18-5 是 W 的精确抽样分布的概率分布图形，表 18-10 中给出 W 的可能值以及对应的精确概率。虽然我们不要求你能得到这个精确的抽样分布，但我们将用它检验两个总体是相同的假设。

如同我们用其他抽样分布一样，让我们用表 18-5 中的 W 的抽样分布计算检验的 p-值，表 18-9 显示 4 名大学生的秩和为 $W=14$，由于这个 W 值位于抽样分布的下侧，我们先计算下侧概率 $P(W\leqslant 14)$，因此，我们有

$$\begin{aligned} P(W \leqslant 14) &= P(10) + P(11) + P(12) + P(13) + P(14) \\ &= 0.0079 + 0.0079 + 0.0159 + 0.0238 + 0.0397 = 0.0952 \end{aligned}$$

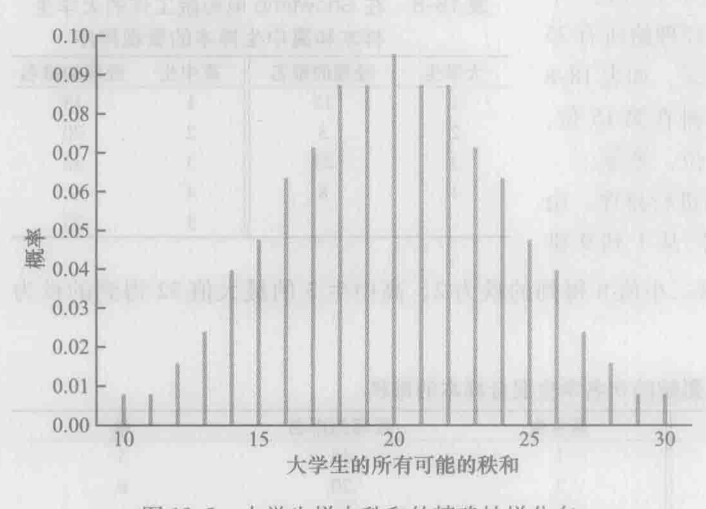


图 18-5 大学生样本秩和的精确抽样分布

表 18-10 大学生样本秩和的精确抽样分布的概率

W	概率	W	概率
10	0.007 9	21	0.087 3
11	0.007 9	22	0.087 3
12	0.015 9	23	0.071 4
13	0.013 8	24	0.063 5
14	0.039 7	25	0.047 6
15	0.047 6	26	0.039 7
16	0.063 5	27	0.0023 8
17	0.071 4	28	0.015 9
18	0.087 3	29	0.007 9
19	0.087 3	30	0.007 9
20	0.095 2		

双侧 p-值 $=2\times 0.0952=0.1904$，在显著性水平 $\alpha=0.05$ 下，由于 p-值 >0.05，因此 MWW 检验的结论是，我们不能拒绝大学生总体与高中生总体是相同的原假设，即 4 名大学生样本和 5 名高中生样本不能提供两个总体存在差异的统计证据。这是对判断提出怀疑的理想时刻，在得出最终结论之前需要考虑用大样本做进一步的研究。

MWW 检验的许多应用涉及较大的样本而非前面出现的例子。对于这些应用，我们使用基于正态分布的 W 抽样分布的一个大样本近似。事实上我们注意到，对于如此少的样本容量 4 和 5，图 18-5 中 W 的抽样分布显示出非常好的正态分布近似。我们将利用前面例子使用的混合样本秩的方法，但我们用正态分布近似而非 W 的精确抽样分布来计算 p-值，并得出结论。

我们通过考察第三国民银行的情况来说明应用正态分布近似进行 MWW 检验。银行经理监控两个分行的支票账户余额，并想确定两个分行账户余额的总体是否相同。银行经理抽取了两个独立的账户余额样本，其中分行 1 的样本容量为 $n_1=12$，分行 2 的样本容量为 $n_2=10$，数据如表 18-11 所示。

表 18-11　第三国民银行两个分行的账户余额

分行 1		分行 2	
账户	余额（美元）	账户	余额（美元）
1	1 095	1	885
2	955	2	850
3	1 200	3	915
4	1 195	4	950
5	925	5	800
6	950	6	750
7	805	7	865
8	945	8	1 000
9	875	9	1 050
10	1 055	10	935
11	1 025		
12	975		

同前面的做法一样，MWW 检验的第一步是将混合数据由小到大进行排序[⊖]。对表 18-11 中的 22 个混合观测值，我们发现最小值（分行 2 的账户 6）为 750 美元，我们把它的秩定为 1；第二小值（分行 2 的账户 5）为 800 美元，我们把它的秩定为 2，第三小值（分行 1 的账户 7）为 805 美元，我们把它的秩定为 3，以此类推。在对混合数据进行排秩时，我们可能会发现有两个或多个数值是相同的，这种情形下，将相等数值在混合数据集中所处位置的平均秩赋予它们。例如，分行 1 的账户 6 和分行 2 的账户 4，它们的余额都是 950 美元。当混合数据从小到大进行排序时，两个 950 美元在混合数据集中的位置是 12 和 13，因此，这两个账户被定为平均秩 $(12+13)/2=12.5$，表 18-12 是混合样本的秩。

表 18-12　混合账户余额样本的秩

分行	账户	余额	秩	分行	账户	余额	秩
2	6	750	1	1	6	950	12.5
2	5	800	2	2	4	950	12.5
1	7	805	3	1	2	955	14
2	2	850	4	1	12	975	15
2	7	865	5	2	8	1 000	16
1	9	875	6	1	11	1 025	17
2	1	885	7	2	9	1 050	18
2	3	915	8	1	10	1 055	19
1	5	925	9	1	1	1 095	20
2	10	935	10	1	4	1 195	21
1	8	945	11	1	3	1 200	22

现在我们回到两个单独的样本，并从表 18-12 中得出每一个账户余额的秩。这些结果列在表 18-13 中。下一步是求出每个样本的秩和：样本 1 为 169.5 和样本 2 为 83.5。正如前文所述，我们遵循使用样本 1 的秩和作为检验统计量 W 的方法，因此，我们得到 $W=169.5$。当两个样本的容量都大于或等于 7 时，W 的抽样分布可以用正态分布近似。在假定原假设为真和总体相同的条件下，检验统计量 W 的抽样分布如下所述。

表 18-13　来自第三国民银行两个样本数据的混合秩

分行 1			分行 2		
账户	余额（美元）	秩	账户	余额（美元）	秩
1	1 095	20	1	885	7
2	955	14	2	850	4
3	1 200	22	3	915	8
4	1 195	21	4	950	12.5
5	925	9	5	800	2
6	950	12.5	6	750	1
7	805	3	7	865	5
8	945	11	8	1 000	16
9	875	6	9	1 050	18
10	1 055	19	10	935	10
11	1 025	17		秩和	83.5
12	975	15			
	秩和	169.5			

⊖ 用手工对混合样本排秩需要一些时间，计算机程序可以用于排秩，既快速又有效率。

相同总体的 W 的抽样分布

$$均值：\mu_W = n_1(n_1 + n_2 + 1)/2 \quad (18\text{-}5)$$

$$标准差：\sigma_W = \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12} \quad (18\text{-}6)$$

分布形式：当 $n_1 \geqslant 7$ 且 $n_2 \geqslant 7$ 时，近似正态分布。

在样本容量 $n_1 = 12$ 和 $n_2 = 10$ 的情况下，利用式（18-5）和式（18-6），得到抽样分布的均值和标准差为

$$\mu_W = 1/2 \times n_1 \times (n_1 + n_2 + 1) = 1/2 \times 12 \times (12 + 10 + 1) = 138$$

$$\sigma_W = \sqrt{1/12 \times n_1 \times n_2 \times (n_1 + n_2 + 1)} = \sqrt{1/12 \times 12 \times 10 \times (12 + 10 + 1)} = 15.1658$$

图 18-6 是 W 的抽样分布所使用的正态分布。

我们继续进行 MWW 检验，并在显著性水平 0.05 下得出结论。由于统计量 W 是离散型的，正态分布是连续型的，我们对正态分布近似再次使用连续性校正因子。由于 $W = 169.5$ 在抽样分布的上侧，我们用如下的 p-值计算：

$$P(W \geqslant 169.5) = P\left(z \geqslant \frac{169 - 138}{15.1658}\right) = P(z \geqslant 2.04)$$

使用标准正态分布表和 $z = 2.04$，双侧 p-值为 $2 \times (1 - 0.9793) = 0.0414$。由于 p-值 $\leqslant 0.05$，所以拒绝 H_0，并得出两个分行的账户余额总体不相同。检验统计量 W 的上侧值表明分行 1 的账户余额总体有较大的倾向。

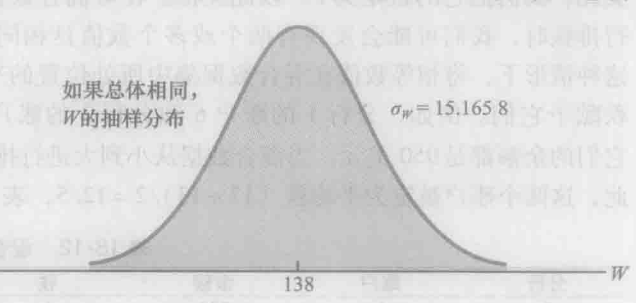


图 18-6 第三国民银行例子中 W 的抽样分布

最后，MWW 检验的一些应用使其适合于假定两个总体具有相同的形态[⊖]，即使总体不同，也仅仅是分布位置上的移动。如果两个总体具有相同的形态，假设检验可以根据两个总体中位数之差来陈述。中位数之间的任何差异可以理解为一个总体与另一个总体相比的位置移动，在这种情况下，关于两个总体中位数的 MWW 检验的三种形式如下：

双侧检验	下侧检验	上侧检验
H_0：中位数$_1$ − 中位数$_2$ = 0	H_0：中位数$_1$ − 中位数$_2$ $\geqslant 0$	H_0：中位数$_1$ − 中位数$_2$ $\leqslant 0$
H_a：中位数$_1$ − 中位数$_2$ $\neq 0$	H_a：中位数$_1$ − 中位数$_2$ < 0	H_a：中位数$_1$ − 中位数$_2$ > 0

注释和评论

附录 18A 中描述了 MWW 检验的 Minitab 程序，Minitab 假定两个总体具有相同的形态。因此，Minitab 描述了关于两个总体中位数的检验结果。如果你认为假定“相同形态”不合适，Minitab 的结果仍然可以应用，但你需要解释检验两个总体相同的原假设的检验结果。

练 习

应用

☆18. 对两种燃料添加剂进行检验以确定它们对汽油行驶里程的影响。对于添加剂 1，我们检测了 7 辆汽车；对于添加剂 2，我们检验了 9 辆汽车。下面的数据是使用两种添加剂所得到的每加仑汽油行驶里程数。在显著性水平 $\alpha = 0.05$ 下，用 MWW 检验，来确定两种添加剂对汽油行驶里程的影响是否存在显著差异。

添加剂 1	添加剂 2
17.3	18.7
18.4	17.8
19.1	21.3
16.7	21.0

⊖ 如果能得出两个总体具有相同的形态的假定，MWW 检验就变成了关于两个总体中位数之差的检验。

（续）

添加剂1	添加剂2
18.2	22.1
18.6	18.7
17.5	19.8
	20.7
	20.2

20. 拥有同等学力的男性和女性的收入差异正在缩小，但是收入并没有达到接近的程度。具有学士学位的7名男性和7名女性收入数据如下表所示（单位：1 000美元）。

男性	35.6	80.5	50.2	67.2	43.2	54.9	60.3
女性	49.5	40.4	32.9	45.5	30.8	52.5	29.8

a. 男性收入的中位数是多少？女性呢？

b. 在显著性水平 $\alpha=0.05$ 下，对总体分布相同的假设进行检验，你的结论如何？

22.《彭博商业周刊》每年公布有关全世界最大的1 000家公司的统计数据。一家公司的市盈率（P/E）是这家公司股票的当前价格除以最近12个月的每股收益。下表中列出了10家日本公司和12家美国公司的市盈率。美国公司和日本公司市盈率之间是否存在显著差异？使用MWW检验和 $\alpha=0.01$ 来支持你的结论。

日本		美国	
公司	市盈率	公司	市盈率
Sumitomo Corp.	153	Gannet	19
Kinden	21	Motorola	24
Heiwa	18	Schlumberger	24
NCP Japan	125	Oracle Systems	43
Suzuki Motor	31	Gap	22
Fuji Bank	213	Winn-Dixie	14
Sumitomo Chemical	64	Ingersoll-Rand	21
Seibu Railway	666	American Electric Power	14
Shiseido	33	Hercules	21
Toho Gas	68	Times Mirror	38
		WellPoint Health	15
		Northern States Power	14

24. 某一特定品牌的微波炉在达拉斯的10家商场和在圣安东尼奥的13家商场中的标价如下。在显著性水平0.05下，检验该微波炉的价格在两个城市之间是否相同（单位：美元）。

达拉斯	圣安东尼奥
445	460
489	451
405	435
485	479
439	475
449	445
436	429
420	434
430	410
405	422
	425
	459
	430

18.4 克鲁斯卡尔－沃利斯检验

在本节我们将假设检验的非参数方法扩展到三个或三个以上总体的情形。当我们使用数量型数据并假定总体服从标准差相等的正态分布时，我们考虑第13章中的参数检验。以每个总体的独立随机样本为依据，我们用 F 分布来检验各总体之间的差异。

非参数的**克鲁斯卡尔－沃利斯检验**（Kruskal-Wallis test）的依据是对来自 k 个总体的 k 个独立随机样本的分析。这个方法既可使用顺序型数据也可使用数量型数据，并且不需要假定总体服从正态分布。原假设和备择假设的一般形式如下：

$$H_0\text{：所有总体是相同的}$$

$$H_a\text{：并非所有总体都是相同的}$$

如果 H_0 被拒绝，我们将得出总体之间存在差异，与其他总体相比较，一个或多个总体提供较小或较大的数值。我们用下面的例子来阐述克鲁斯卡尔－沃利斯检验。

威廉姆斯制造公司从3所不同的大学招聘管理工作人员。最近该公司的人事主管开始审核管理人员的年度业绩报告，试图确定毕业于这3所大学的管理人员的业绩评分是否存在差异。大学A毕业的7名管理人员，大学B

毕业的6名管理人员，大学C毕业的7管理人员的三个独立样本的业绩分数据如表18-14所示。每名管理人员的总体业绩分在0~100，其中100是最高的业绩分。假设我们想检验，3个管理人员总体的业绩分是否存在差异，我们使用的显著性水平为0.05。

表18-14 20名威廉姆斯员工的业绩分

大学A	大学B	大学C
25	60	50
70	20	70
60	30	60
85	15	80
95	40	90
90	35	70
80		75

克鲁斯卡尔-沃利斯检验的第1步是对混合样本数据从低到高进行排序，利用表18-14中的20个观测值，大学B的第4名管理人员的最低业绩分15得到的秩为1，大学A的第5名管理人员的最高业绩分95得到的秩为20，表18-15显示了业绩分数和它们相应的秩。注意，并列的业绩分60，70，80和90，我们定义其秩为平均秩，表18-15还给出了3个样本中每一个的秩和。

表18-15 3个样本的混合秩

大学A	秩	大学B	秩	大学C	秩
25	3	60	9	50	7
70	12	20	2	70	12
60	9	30	4	60	9
85	17	15	1	80	15.5
95	20	40	6	90	18.5
90	18.5	35	5	70	12
80	15.5		秩和27	75	14
	秩和95				秩和88

克鲁斯卡尔-沃利斯检验的统计量利用3个样本的秩和，其计算公式如下。

克鲁斯卡尔-沃利斯检验的统计量

$$H=\left[\frac{12}{n_T(n_T+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right]-3(n_T+1) \tag{18-7}$$

式中，k代表总体的个数；n_i代表样本i中观测值的个数；$n_T=\sum_{i=1}^{k}n_i$，代表所有样本的观测值总数；R_i为样本i的秩和。

克鲁斯卡尔和沃利斯已经证明，在各个总体相同的原假设下，H的抽样分布可以近似服从自由度为$k-1$的χ^2分布。如果k个总体的每个样本的容量都大于或等于5，则这种近似是可以被普遍接受的。如果这个统计量H的值非常大，则总体相同的原假设将被拒绝。因此，克鲁斯卡尔-沃利斯检验总是表示成一个上侧检验。对表18-15中的样本数据，检验统计量的计算结果如下：

样本容量为

$$n_1=7 \qquad n_2=6 \qquad n_3=7$$

并且

$$n_T=\sum_{i=1}^{3}n_i=7+6+7=20$$

利用每一个样本的秩和，克鲁斯卡尔和沃利斯检验的统计量为

$$H=\left[\frac{12}{n_T(n_T+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right]-3(n_T+1)=\frac{12}{20\times 21}\left[\frac{95^2}{7}+\frac{27^2}{6}+\frac{88^2}{7}\right]-3\times(20+1)=8.92$$

我们现在利用χ^2分布表（附录B中的表3）来确定检验的p-值。根据自由度为$k-1=3-1=2$，我们看到$\chi^2=7.378$对应的χ^2分布的上侧面积为0.025，而$\chi^2=9.21$对应的χ^2分布的上侧面积为0.01。由于$H=8.92$介于7.378和9.21之间，因此，我们可以得出χ^2分布上侧面积在0.01和0.025之间。由于这是上侧检验，我们可以得到p-值在0.01和0.025之间。利用Minitab或Excel得出$\chi^2=8.92$的精确p-值为0.0116。由于p-值$\leqslant\alpha=0.05$，我们拒绝H_0，并得出3个总体不全相同的结论。毕业于3所大学的管理人员业绩分的3个总体是不同的，并且差

异显著地依赖于大学。由于毕业于大学B的管理人员的样本秩和相对较低，所以公司有理由减少从大学B招聘员工，或者在作出录用决定之前至少需要对毕业于大学B的毕业生进行更加深入的考核。

最后，我们注意到在克鲁斯卡尔－沃利斯检验的一些应用中，作出总体有相同的形态的假定可能是合适的，如果有所不同，也仅仅是一个或多个总体在位置上的移动。如果假定 k 个总体具有相同的形态，假设检验可以根据总体中位数来表述。[⊖]因此，克鲁斯卡尔－沃利斯检验的假设可写成如下形式：

$$H_0: 中位数_1 = 中位数_2 = \cdots = 中位数_k$$

$$H_a: 并非所有的中位数都相等$$

注释和评论

1. 本节例子中使用员工业绩分的数量型数据进行克鲁斯卡尔－沃利斯检验。如果20名员工的业绩分数据是顺序等级，这个检验也能使用。在这种情况下，可以直接应用原始数据，将数量型数据转变为秩－顺序数据这一步骤就没有必要了，练习第30题说明了此情形。
2. 附录18A中描述了克鲁斯卡尔－沃利斯检验的Minitab步骤。Minitab假定总体具有相同的形态，因此，Minitab描述的克鲁斯卡尔－沃利斯检验可作为总体中位数之间差异的检验。如果你认为假定“相同形态”不合适，你仍然可以应用Minitab，但你需要解释检验所有总体相同的原假设的检验结果。

练习

应用

☆26. 由15名消费者组成的一个样本对下面三种不同产品给出了以下评分。5名消费者被随机指定去检验和评价一种产品。利用克鲁斯卡尔－沃利斯检验和显著性水平 $\alpha=0.05$ 来确定这三种产品的评分之间是否存在显著差异。

产品		
A	B	C
50	80	60
62	95	45
75	98	30
48	87	58
65	90	57

☆28. 每周进行三天、每次40分钟的下列运动之一，可以导致体重下降。下面的样本数据为三种不同的运动在40分钟内所消耗的卡路里数据。这些数据是否表明三种运动所消耗的卡路里数量是不同的？在显著性水平0.05下，你的结论如何？

游泳	网球	自行车
408	415	385
380	485	250
425	450	295
400	420	402
427	530	268

30. 一家大公司将其许多第一流的经理送去管理技能培训班学习。有四家不同的管理发展中心提供这种培训班。人力资源总监想要了解这四家中心所提供的培训质量是否不同。从每一家培训中心中随机抽取5名员工组成样本，然后将这些员工按管理技能进行排秩。秩1表明员工有最好的管理技能，排秩如下所示。利用 $\alpha=0.05$ 检验这四家中心所提供的培训是否存在显著不同。

课程			
A	B	C	D
3	2	19	20
14	7	16	4
10	1	9	15
12	5	18	6
13	11	17	8

18.5 秩相关

在第3章介绍的皮尔逊积矩相关系数是两个数量型数据变量之间线性关系的度量。在本节，当使用顺序或排

⊖ 如果能得出总体具有相同的形态的假定，克鲁斯卡尔－沃利斯检验就变成了关于 k 个总体中位数的检验。

秩数据时，我们给出两个变量之间联系的相关性度量。**斯皮尔曼秩相关系数**正是为这一目的而提出来的。

斯皮尔曼秩相关系数

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{18-8}$$

式中，n 为样本中观测值的个数；x_i 为对于第 1 个变量的第 i 观测值的秩；y_i 为对于第 2 个变量的第 i 观测值的秩；$d_i = x_i - y_i$。

让我们说明斯皮尔曼秩相关系数的应用。一家公司想确定那些在雇用时预期会成为较好推销员的人是否能够真正拥有良好的销售业绩。为了研究这个问题，人事主管仔细回顾了现有销售队伍中 10 名成员原来的求职面试报告、学习记录和推荐信。审核之后，这位主管根据这 10 个人在聘用期成功的潜力进行排序，最有潜力的秩为 1。然后又收集每一位推销员聘用前两年的实际销售业绩。根据每一位推销员的实际销售业绩，得到 10 名推销员实际销售业绩的第二次排秩。表 18-16 给出了以潜力和实际销售业绩为依据的排秩。

表 18-16 销售潜力与两年实际销售数据

推销员	潜力排秩	两年销售量（个）	两年销售量排秩
A	2	400	1
B	4	360	3
C	7	300	5
D	1	295	6
E	6	280	7
F	3	350	4
G	10	200	10
H	9	260	8
I	8	220	9
J	5	385	2

让我们来计算表 18-16 中数据的斯皮尔曼秩相关系数。计算过程汇总在表 18-17 中。我们首先计算每一位销售员两个秩之间的差 d_i，列在第 4 列，d_i^2 的和为 44 列在第 5 列。这个值和样本容量 $n=10$ 用于计算秩相关系数 $r_s=0.73$，如表 18-17 所示。

斯皮尔曼秩相关系数的取值范围是 $-1.0 \sim +1.0$，而且它的解释与数量型数据的皮尔逊积矩相关系数相似，秩相关系数接近于 $+1.0$ 意味着两个变量之间有很强的正相关关系；而秩相关系数接近于 -1.0 意味着两个变量之间存在很强的负相关关系；秩相关系数为 0 意味着两个变量之间没有相关关系。在本例中，$r_s=0.733$ 表明销售潜力的秩和实际销售业绩的秩之间存在正相关关系。在聘用期间潜力的秩较高的推销员在前两年实际销售业绩的排秩也比较高。

表 18-17 销售潜力与实际销售业绩的斯皮尔曼秩相关系数的计算表

推销员	x_i = 销售潜力排秩	y_i = 销售业绩排秩	$d_i = x_i - y_i$	d_i^2
A	2	1	1	1
B	4	3	1	1
C	7	5	2	4
D	1	6	5	25
E	6	7	1	1
F	3	4	1	1
G	10	10	0	0
H	9	8	1	1
I	8	9	1	1
J	5	2	3	9
				$\sum d_i^2 = 44$

$$r_s = 1 - \frac{6\sum d_i^2}{n\ (n^2-1)} = 1 - \frac{6\times 44}{10\times\ (100-1)} = 0.733$$

此时，我们可能希望利用样本秩相关系数 r_s 来推断总体的秩相关系数 ρ_s。为此，我们检验下面的假设。

$$H_0: \rho_s = 0$$
$$H_a: \rho_s \neq 0$$

在假定原假设为真和总体的秩相关系数为 0 的条件下，用来进行检验的检验统计量 r_s 的抽样分布如下。

r_s 的抽样分布

$$均值:\mu_{r_s} = 0 \tag{18-9}$$

$$标准差:\sigma_{r_s} = \sqrt{\frac{1}{n-1}} \tag{18-10}$$

分布形式：$n \geqslant 10$ 时，近似服从正态分布。

销售潜力与销售业绩的样本秩相关系数为 $r_s=0.733$。由式（18-9）我们有 $\mu_{r_s}=0$，由式（18-10）我们有 $\sigma_{r_s}=\sqrt{1/(10-1)}=0.333$。由于 r_s 的抽样分布近似服从正态分布，标准正态随机变量 z 成为检验的统计量，而且

$$z = \frac{r_s - \mu_{r_s}}{\sigma_{r_s}} = \frac{0.733-0}{0.333} = 2.20$$

利用标准正态分布表和 $z=2.20$，我们看到双侧 p-值为 $2\times(1-0.9861)=0.0278$。在显著性水平 0.05 下，由于 p-值 $\leqslant \alpha=0.05$，因此总体秩相关系数为 0 的原假设被拒绝。检验结果表明聘用期间的销售潜力与实际销售业绩之间存在显著的秩相关关系。

注释和评论

斯皮尔曼秩相关系数与应用于秩－等级数据的皮尔逊积矩相关系数是相同的。在附录 18A 和 18B 中，我们演示了如何利用皮尔逊积矩相关系数的 Minitab 和 Excel 的相关系数工具来计算斯皮尔曼秩相关系数。

练习

方法

☆32. 考虑下列 10 个元素所组成的样本的排秩数据。

元素	x_i	y_i	元素	x_i	y_i
1	10	8	6	2	7
2	6	4	7	8	6
3	7	10	8	5	3
4	3	2	9	1	1
5	4	5	10	9	9

a. 计算数据的斯皮尔曼秩相关系数。

b. 在显著性水平 $\alpha=0.05$ 下，检验秩相关的显著性。你的结论如何？

应用

☆34. 下面是 11 个州以每名学生费用（1 为最高，11 为最低）和师生比（1 为最低，11 为最高）为依据的的排秩数据。

州	每名学生费用	师生比
亚利桑那	9	10
科罗拉多	5	8
佛罗里达	4	6
爱达荷	2	11
艾奥瓦	6	4
路易斯安那	11	3
马萨诸塞	1	1
内布拉斯加	7	2

（续）

州	每名学生费用	师生比
北达科他	8	7
南达科他	10	5
华盛顿	3	9

a. 每名学生费用与师生比之间的秩相关系数是多少？请讨论。

b. 在显著性水平 $\alpha=0.05$ 下，每名学生费用与师生比之间是否存在显著相关关系？

36. 一个职业高尔夫球员样本的击球距离和击球杆数的排秩结果如下。这些高尔夫球员的击球距离和击球杆数之间的秩相关系数是多少？在显著性水平 0.10 下，检验相关系数的显著性。

职业高尔夫选手	击球距离	击球杆数
Fred Couples	1	5
David Duval	5	6
Ernie Els	4	10
Nick Faldo	9	2
Tom Lehman	6	7
Justin Leonard	10	3
Davis Love Ⅲ	2	8
Phil Mickelson	3	9
Greg Norman	7	4
Mark O' Meara	8	1

小结

本章我们给出了几种被归类为非参数方法的统计程序。因为非参数方法不仅可以应用于数量型数据，而且可以应用于分类型数据，同时也不要求对总体分布的假设，所以它们扩展了统计分析研究对象问题的范围。

符号检验是检验一个总体中位数或匹配样本的非参数方法。数据必须汇总成两个类别，其中一个记为加号，另一个记为减号。当可以用的是数量型数据时，威尔科克森符号秩检验分析来自两个总体的匹配样本。它对配对观测值之差的分布没有假设，只需要假定配对观测值之差是对称的。威尔科克森符号秩检验用于确定配对总体中位数之差是否为零。该检验也可以用于对对称总体中位数进行推断。

曼－惠特尼－威尔科克森检验是一种基于两个独立随机样本，用于检验两个总体间差异的非参数方法。它是和两个总体均值之差的参数 t 检验相对应的方法。从两个混合样本的数据中得到混合秩，用第1个样本的秩和作为MWW检验的统计量。在许多应用中，样本容量充分大，足以用带有连续性校正因子的正态分布近似来进行假设检验。如果对总体的分布没有假定，MWW方法检验两个总体是否相同。如果假定两个总体具有相同的形态，检验给出了关于两个总体中位数之间差异的推断。

克鲁斯卡尔－沃利斯检验将MWW检验扩展到三个或多个的总体，它是和检验三个或多个正态总体均值间差异的ANOVA参数检验相对应的非参数方法。克鲁斯卡尔－沃利斯检验不需要对总体的分布有任何假定，并且使用 k 个总体相同的原假设。如果假定总体具有相同的形态，检验给出了关于 k 个总体中位数之间差异的推断。在本章的最后一节中，我们介绍了斯皮尔曼秩相关系数，它以秩－顺序数据为依据的两个变量相关性的一种度量。

关键术语

parametric methods 参数方法 通常从总体概率分布服从正态分布的假定开始的统计方法，然后推导出检验统计量的抽样分布，并用来对总体的一个或多个参数（如总体均值 μ 或总体标准差 σ）进行推断。

nonparametric methods 非参数方法 对总体的概率分布的形式没有假设要求的统计方法，通常也称为无分布方法。一些方法既可以使用分类型数据也可以使用数量型数据。

distribution-free method 无分布方法 对总体的概率分布没有假定的统计方法。

sign test 符号检验 对总体中位数的假设，或为了确定基于匹配样本的两个总体之间差异的一种非参数统计检验方法。数据汇总成两个类别，其中一个记为加号，另一个记为减号，$p=0.50$ 的二项分布是检验统计量的抽样分布。

Wilcoxon signed-rank test 威尔科克森符号秩检验 以匹配样本为依据的两个总体中位数之差的一种非参数检验方法。该方法使用数量型数据，并假定配对观测值之差的分布是对称的。配对观测值之差的数据可以用于对两个总体的中位数进行推断，该检验也可以对一个对称总体的中位数进行推断。

Mann-Whitney-Wilcoxon（MWW）test 曼－惠特尼－威尔科克森（MWW）检验 以来自每个总体的独立样本为依据的两个总体之间差异的一种非参数统计检验方法，原假设是两个总体相同。如果假定总体具有相同的形态，检验给出了关于两个总体中位数之间差异的推断。

Kruskal-Wallis test 克鲁斯卡尔－沃利斯检验 以来自每个总体独立样本的分析为依据的三个或三个以上总体之间差异的一种非参数检验方法，原假设为所有总体相同。如果假定总体具有相同的形态，检验给出了关于多个总体中位数之间差异的推断。

Spearman rank-correlation Coefficient 斯皮尔曼秩相关系数 以秩－顺序数据为依据的两个变量之间联系的相关性度量。

重要公式

符号检验：正态近似

$$\text{均值}：\mu = 0.50n \tag{18-1}$$

$$\text{标准差}：\sigma = \sqrt{0.25n} \tag{18-2}$$

威尔科克森符号秩检验：正态近似

$$\text{均值}：\mu_{T^+} = \frac{n(n+1)}{4} \tag{18-3}$$

标准差：$\sigma_T = \sqrt{\dfrac{n(n+1)(2n+1)}{24}}$ (18-4)

曼－惠特尼－威尔科克森检验：正态近似

均值：$\mu_W = n_1(n_1 + n_2 + 1)/2$ (18-5)

标准差：$\sigma_W = \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}$ (18-6)

克鲁斯卡尔－沃利斯检验统计量

$$H = \left[\frac{12}{n_T(n_T+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(n_T + 1) \quad (18\text{-}7)$$

斯皮尔曼秩相关系数

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (18\text{-}8)$$

补充练习

38. 一项调查询问了如下问题：你赞同还是反对给予将子女送入私立学校的家庭提供来自国家税收的代金券或者减税？在接受调查的2 010人中，有905人赞同进行这种资助，有1 045人反对进行这种资助，还有60人没有表明观点。在显著性水平0.05下，这些数据是否能够表明人们对将子女送入私立学校的家庭给予资助存在分歧？

40. 请来12名家庭主妇估计两种型号冰箱的零售价格。下面是她们估计的零售价格。利用这些数据和显著性水平0.05，就家庭主妇对零售价格的看法，确定两种型号冰箱之间是否存在差异。

家庭主妇	型号1（美元）	型号2（美元）
1	850	1 100
2	960	920
3	940	890
4	900	1 050
5	790	1 120
6	820	1 000
7	900	1 090
8	890	1 120
9	1 100	1 200
10	700	890
11	810	900
12	920	900

42. 以下的数据是两条生产线上所生产的产品重量。在显著性水平$\alpha = 0.05$下，检验两条生产线所生产的产品重量是否存在差异。

生产线1	生产线2	生产线1	生产线2
13.6	13.7	13.3	14.4
13.8	14.1	13.6	14.8
14.0	14.2	12.9	14.5
13.9	14.0	14.4	14.3
13.4	14.6		15.0
13.2	13.5		14.9

44. 由20名已在一家公司工作了三年的工程师组成样本，对这些工程师按管理潜力进行排秩。其中有一些工程师已经参加过公司的管理培训课程，另一些人参加了当地大学的短期管理培训课程，剩下的则没有参加过任何培训课程。利用下面的排秩数据，在显著性水平$\alpha = 0.025$下，检验三组工程师的管理潜力的显著差异。

未参加培训	公司课程培训	短期课程培训
16	12	7
9	20	1
10	17	4
15	19	2
11	6	3
13	18	8
	14	5

46. 下面是15名学生统计学课程期中和期末考试成绩的排秩样本数据。计算这些数据的斯皮尔曼秩相关系数，并在显著性水平$\alpha = 0.10$下，检验显著相关性。p-值是多少？你的结论如何？

秩		秩	
期中	期末	期中	期末
1	4	9	6
2	7	10	9
3	1	11	14
4	3	12	15
5	8	13	11
6	2	14	10
7	5	15	13
8	12		

第 19 章

质量管理的统计方法

CHAPTER 19

实践中的统计

陶氏化学公司[一]

得克萨斯州，弗里波特

1940年，陶氏化学公司（Dow Chemical Company）在得克萨斯州墨西哥湾沿岸购买了800英亩的土地，用以建造一个镁生产设施。如今那片厂址已经扩展到超过5 000英亩，该公司也成为世界上最大的石油化工复合型企业之一。得克萨斯州陶氏化学公司的产品包括镁、苯乙烯、塑料、胶粘剂、溶剂、乙二醇和氯气等。有些产品制造出来仅仅是为了用于其他产品的生产，但是很多产品则是成为医药、牙膏、狗粮、水管、冰箱、奶桶、垃圾袋、洗发水和家具等产品的必要的基础性原料。

得克萨斯州陶氏化学公司所生产的镁产量占全世界总产量的30%以上。镁是一种极轻的金属，它用于生产网球拍、衣物箱以及镁合金轮胎等各类的产品。该公司的镁技术部门是首先利用统计质量管理方法培训其技术人员和管理人员的部门，统计质量管理的最初的一些成功应用出现在化学处理过程之中。

干燥剂的生产是其中的一个应用。在这个应用中，他们每隔一定周期抽取产品组成样本，计算每个样本的平均值，并且将它们记录在$\bar{x}$控制图上。利用这种控制图，陶氏化学公司的分析家可以监控产品的生产过程，随时预警过程的非正常运转。一个实例是，分析家开始观察样本均值的数值，该数值显示有一次过程的运转在它的设计范围之外。经过对控制图和运转情况的进一步观察，分析家发现差异可以追溯到某个操作人员。在该操作人员重新培训后，通过记录的$\bar{x}$控制图，可以看到过程的质量已经有了重大的改进。

陶氏化学公司在任何使用统计质量管理的地方都可以取得质量的大幅度改进。公司每年节约成本数十万美元，并且不断地发现新的应用。

本章我们将阐述陶氏化学公司所使用的$\bar{x}$控制图是如何构造的。控制图是统计质量控制的一部分，它被称为统计过程控制。我们也将讨论以样本为根据来决定接收或者拒绝一组项目的质量管理方法。

美国质量协会[二]（American Society for Quality，ASQ）对质量是这样定义的："质量是产品或服务的所有性质和特征，这些性质和特征使得该产品或服务能够满足特定的需要。"换句话说，质量可以测量产品或服务满足顾客需要的程度。很多企业意识到在全球经济竞争的今天，它们必须追求高水平的质量。因此，对质量的检验和维护有必要在方法上加以重视。

今天，作为高绩效组织基础的顾客驱动视角改变了质量问题所包含的范围，从简单地排除生产线上的缺陷产品到建立运用广泛的公司战略。扩大质量的范围自然导致**全面质量**（total quality，TQ）的概念。

全面质量是以人为本的管理系统，其目的在于不断降低实际成本以持续提升顾客满意度。全面管理是一种全面系统的方法（不是一个单独的领域或工作项目）和高水平战略的组成部分；它横跨职能与部门运转，从上到下包括全体员工，通过前向后向延展将质量管理扩展到供应链与消费链领域。全面质量强调学习和适应并将持续改变作为组织成功的关键。[三]

不管全面质量在不同组织中是如何实现的，它主要依据三个基本原理：以顾客和股东为中心；将参与精神与团队合作贯穿整个组织；以不断改进和学习为中心。在本章第一节，我们简要介绍质量管理的三个框架：马尔科姆·鲍德里奇全国质量奖、ISO 9000标准和六西格玛原则。在后两节中，我们介绍两种用于检测过程的统计工具：统计过程控制和接受抽样。

[一] 作者感谢为"实践中的统计"提供了本案例的陶氏化学公司镁技术部门经理Clifford B. Wilson先生。

[二] 美国质量协会的愿景："通过制定作为全局的重点、组织的责任和个人的道德规范的质量标准，美国质量协会成为那些探求质量观念、技术和工具来改变他们自己和他们世界的每一个人的团体"（ASQ website）。

[三] J. R. Evans，W. M. Lindsay. The Management for Quality and Performance Excellent，8th ed.（Cincinnati，Oh：South-Western，2011），p. 11。

19.1 理念和框架

在20世纪初期，质量管理实践仅限定于检验已完成的产品并剔除有缺陷的项目。但是这一切的改变归功于一位名为沃尔特A. 休哈特（Walter A. Shewhart）的年轻工程师的开拓性成就。1917年从哥伦比亚大学物理系获得博士学位后，休哈特博士加入西部电话公司，在工程检验部工作。1924年休哈特博士编写了今天众所周知的过程控制基本原理的一套准则备忘录，同时他的备忘录中还包含被公认为统计控制图的图表。休哈特博士在贝尔电话实验室继续从事他的质量工作直到1956年退休，他将统计学、工程学、经济学等学科融合在一起，从而改变了工业历史的进程。休哈特博士是公认的统计质量控制之父，也是美国质量管理协会的首批名誉会员。

还有两个人对质量有重大影响，其中一人是休哈特博士的学生之一W. 爱德华·戴明（W. Edwards Deming）博士㊀，另一人是约瑟夫·朱兰（Joseph Juran）博士。在第二次世界大战后不久，他们在质量管理方面帮助教育日本人。尽管质量是每个人的工作，但是戴明博士强调，注重质量必须由管理者领导。他提出了对管理者至关重要的14个职责要点。例如，戴明认为管理者必须结束对大量检验的依赖；必须终止仅凭价格就做出商业行为的做法；必须寻求在所有生产过程和服务中的不断改进；必须培养团队合作精神的环境；必须消除数量化的目标、口号和确定数量定额的工作标准。也许最重要的是管理者必须创造一个工作环境，在这个环境中，质量和生产率的标准在任何时候都能得到维持。

朱兰为质量下了一个简单的定义：质量必须与实用性相匹配。朱兰的质量方法关注三个质量过程：质量计划、质量控制和质量改进。与戴明的需要在组织中开展重大文化变革的理念相比，朱兰的方案旨在当前的组织系统内提高工作质量。虽然如此，但是在关注高级管理人员的需要，强调持续改进与培训的重要性以及质量控制技术的使用等方面，两种基本原理是相似的。

在质量改进运动中，许多学者发挥了重要作用，包括菲利普B. 克劳士比、A.V. 费根鲍姆、石川馨、田口玄一等人。专门关注质量的专业教材详细描述了他们的贡献。涉及质量活动的所有人的贡献定义了最佳实践，并导致了众多奖项和认证项目的问世。其中两个最重要的项目是美国的马尔科姆·鲍德里奇国家质量奖和ISO 9000国际认证过程。近年来，六西格玛（一种基于严格数据收集和统计分析来改进组织业绩的技术）的使用也在增加。

19.1.1 马尔科姆·鲍德里奇国家质量奖

由美国总统颁发的马尔科姆·鲍德里奇国家质量奖授予在七个方面表现杰出的组织：①领导；②战略策划；③客户和市场；④测量、分析和信息管理；⑤人类资源；⑥流程管理；⑦经营成果。美国国会于1987年通过了奖励计划，以表彰在质量方面做出成就的美国组织，以此提高人们把质量作为竞争优势重要性的意识。该奖项以美国商务部㊁前部长马尔科姆·鲍德里奇命名，他于1981年任职，直至1987年去世。

自1988年首次颁奖以来，马尔科姆·鲍德里奇国家质量奖（BNQP）在深度和影响力方面逐渐拓展。1988年以来，已经分发了近200万册准则手册，通过组织的大规模复制以及电子化途径，其数量得到有效的增长。在连续8年中，假设股票指数（由公开交易的获得鲍德里奇奖的美国公司构成）优于标准普尔500。在某一年，鲍德里奇指数㊂以4.4:1优于标准普尔500。摩托罗拉公司执行副总裁鲍勃·巴内特说："我们申请这个奖，不是为了获奖，而是为了得到鲍德里奇主考官的评价。这个评价是全面的、专业的和富有洞察力的……今天，无论在世界何地，它都是最有经济效益的、有价值的商业咨询活动。"

19.1.2 ISO 9000

ISO 9000是由5个国际标准组成的一个系列标准，由设在瑞士日内瓦的国际标准化组织（ISO）于1987年颁

㊀ 第二次世界大战之后，W. 爱德华·戴明博士成为日本工业顾问。他说服日本高层管理者使用统计质量控制方法。
㊁ 美国商务部的国家标准和技术研究院（NIST）管理马尔科姆·鲍德里奇国家质量项目，可以从NIST的网站上得到更多信息。
㊂ 2004年是鲍德里奇股票研究的最后一年，因为非营利或私有企业接受者的数量在增加。

布。公司可以利用这些标准帮助确定，维持一个有效的质量一致性体系需要有哪些方面。例如，标准描述一个有效质量体系的需要；对确保定期测量和检测设备使其符合标准的需要，以及对维持一个令人满意地记录保持系统的需要。ISO 9000 认证确定一个公司是否遵守其质量体系[一]。总之，ISO 9000 认证覆盖的鲍德里奇质量奖标准不到10%。

19.1.3　六西格玛

在20世纪80年代后期，摩托罗拉认识到改进其产品和服务质量的需要。他们的目标是达到非常高的质量水平，以至于每百万产品出现缺陷的机会不超过3.4。这个质量水平被称为六西格玛质量水平，达到这个质量目标创造的方法被称为**六西格玛**（six sigma）。

一个组织可以承担两类六西格玛方案：

- DMAIC（Define，Measure，Analyze，Improve and Control，界定、测量、分析、改进和控制）帮助重新设计已存在的过程。
- DFSS（六西格玛设计）来于设计新产品、过程或服务。

重新设计已存在的过程和新过程中，六西格玛着重强调统计分析和仔细测量。今天，在帮助组织的经营业绩和过程质量达到鲍德里奇的水平方面，六西格玛已经成为一个主要工具。许多鲍德里奇的主考官视六西格玛为完成鲍德里奇改进方案的理想方法。

六西格玛线和每百万个机会中的缺陷数　在六西格玛术语中，缺陷是指传递给消费者的任何错误或误差。六西格玛过程定义质量性能为百万个机会中的缺陷数（dpmo）。正如我们前面指出的，六西格玛要求的质量水平不超过3.4个dpmo。为了阐明如何测量这个质量水平，我们考虑KJW包装公司的情形。

KJW公司经营一条填装谷物箱的生产线。填装过程的均值为$\mu=16.05$盎司，标准差为$\sigma=0.10$盎司。另外，假设填装重量服从正态分布，填装重量的分布如图19-1所示。假设管理人员认为15.45~16.65盎司是填装过程可以接受的质量限。因此，小于15.45盎司或大于16.65盎司的谷物箱则被认为是有缺陷的。利用Excel[二]，可以看到99.999 999 8%的谷物箱，其填装重量在$16.05-6\times0.10=15.45$（盎司）和$16.05+6\times0.10=16.65$（盎司）之间，换句话说，只有0.000 000 2%的谷物箱填装重量小于15.45盎司或大于16.65盎司。因此在填装过程中发现有缺陷的谷物箱极为罕见，因为在1 000万箱中平均只有2箱存在缺陷。

摩托罗拉早期在六西格玛上的工作使它们确信，过程均值平均而言可以移动高达1.5个标准差。例如，假设KJW的过程均值增加1.5个标准差或$1.5\times0.10=0.15$（盎司），根据这个移动，填装重量正态分布的中心将变为$\mu=16.05+0.15=16.20$（盎司）。因为过程均值$\mu=16.05$盎司，得到谷物箱超过16.65盎司的概率非常小，但是，如果过程均值移动到$\mu=16.20$，这个概率将如何变化？图19-2中显示了这种情况，上质量限16.65在新的过程均值$\mu=16.20$盎司右侧的4.5个标准差。利用这个均值以及Excel，我们发现，一箱重量大于16.65盎司箱的概率为0.000 003 4[三]。因此，如果过程均值向上移动1.5个标准差，则大约有$1\,000\,000\times0.000\,003\,4=3.4$（个）谷物箱超过上线16.65盎司。在六西格玛术语中，我们说过程的质量水平是每百万缺陷机会为3.4。如果KJW的管理人员认为15.45~16.65盎司对填装过程是可以接受的质量限，则KJW填装过程被认为是一个六西格玛过程。于是，如果过程均值停留在目标值$\mu=16.05$盎司的1.5个标准差之内，则可以期待每百万填谷物箱中最多只有3.4箱存在缺陷。

要达到和维持六西格玛质量水平，组织必须强调监控和维持质量。质量保证是指为了达到和维持质量，由某个组织制定的政策、步骤和准则的整套系统。质量保证有两个重要功能：质量工程和质量控制。质量工程的目标包括产品设计和生产过程设计的质量，还包括在生产之前确定质量问题。**质量控制**（quality control）包括一系列的检验和测量以确定是否满足质量标准。如果没有满足质量标准，则可以通过修正或预防措施来达到和维持质量

[一] ISO 9000标准周期性进行修订，以改进标准的质量。

[二] 利用Excel，NORM. S. DIST（6，TRUE）-NORM. S. DIST（-6，TRUE）=0.999 999 998。

[三] 利用Excel，1-NORM. S. DIST（4.5，TRUE）=0.000 003 4。

的一致性。在下面两节中，我们将介绍在质量控制中所使用的两种统计方法。第一种方法是统计过程控制，利用控制图的图形显示来监控生产过程；其目的是确定过程能否继续或者是否应采取修正措施，以达到预期的质量水平。第二种方法是接受抽样，在必须根据样本中发现的质量问题，决定接受或拒绝一组项目时，使用这种方法。

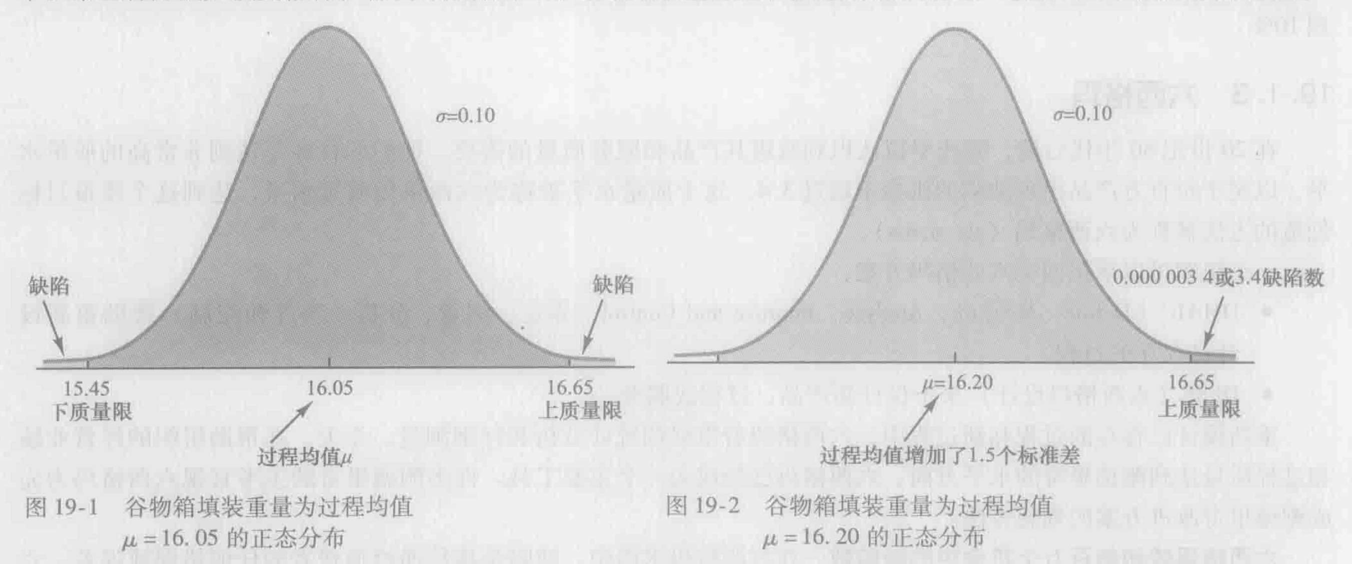


图 19-1 谷物箱填装重量为过程均值 $\mu=16.05$ 的正态分布

图 19-2 谷物箱填装重量为过程均值 $\mu=16.20$ 的正态分布

19.1.4 服务业中的质量

虽然质量控制的根本在制造业，但是对主要重点是提供服务的商业，质量控制也是非常重要的。提供服务的商业例子主要包括保健提供者、法律事务所、旅店、航空公司、饭店和银行。主要提供服务的商业是美国经济的一个非常重要的组成部分。事实上，美国大量的非农业雇员从事提供服务工作。

服务业中的质量效果不是关注测量生产过程的缺陷，而是关注确保消费者满意度和提升消费者体验。通常，由于留住一个消费者比获得一个新消费者的成本要低，因此，设计提升消费者服务的质量控制过程对服务业至关重要。消费者满意度是任何服务型企业成功的关键。

服务业与制造业有很大的不同，而且如何测量和获得质量也有影响。提供的服务常常是无形的（如宿舍顾问的建议）。由于消费者满意度非常主观，因此测量服务中的质量是很有挑战性的。然而，通过测量诸如提供服务的及时性以及进行消费者满意度调查，可以监控质量。这是一些干洗店保证一小时的服务以及汽车服务中心、航空公司和饭店请求你填写有关你服务体验的调查问卷的原因。这也是企业使用顾客积分卡的原因。通过追踪你的购买行为，他们可以更好地了解他们顾客的欲望和需求，随之提供更好的服务。

19.2 统计过程控制

本节我们通过一个连续不断制造货物的生产过程来考虑质量控制程序。以产品生产量的抽样和检验为依据，可以作出是继续生产过程还是调整生产过程的决定，使得项目或正在生产的商品达到可接受的质量标准。⊖

尽管在制造和生产经营中会执行严格的高标准，但是机械设备会不可避免地磨损，振动将会使机械设备的装置难以调整，购买的原材料可能有缺陷，操作人员也可能有失误。所有这些因素可能导致低质量产品的产生。幸运的是，监控生产的程序是行之有效的，能比较早地检测出低质量的产品，生产过程因而得到调整或修正。

如果生产的产品在质量上的变异来源于**可指出的原因**（assignable causes），比如工具的磨损、错误的机器安装、劣质原材料或者操作人员的失误等，则应该立即调整或修正生产过程。另外，如果变异来源于所谓的**一般原因**（common causes），比如在原料、温度和湿度等方面随机出现的差异，生产者可能无法控制，生产过程也不需

⊖ 持续改进是全面质量管理活动的重要概念之一，控制图的最重要的应用是改进过程。

要调整。统计质量控制的主要目标是确定生产产品的质量变异是来源于可指出的原因还是一般原因。

无论何时我们发现可指出的原因，我们的结论是，生产过程处于失控状态。在这种情况下，应该采取纠正措施将生产过程重新调整到可以接受的质量水平。但是，如果生产过程中的产品变异仅仅来源于一般原因，我们的结论是，生产过程处于统计控制状态，或简单地说处于在控状态。在这种情况下，我们没有必要进行改变或者调整[⊖]。

过程控制统计程序的依据是第 9 章介绍的假设检验方法。原假设 H_0 表示生产过程处于在控状态，备择假设 H_a 表示生产过程处于失控状态。表 19-1 表明如何对继续在控状态过程和调整失控状态的过程，作出正确的决定。但是，同其他假设检验程序一样，第一类错误（调整一个在控状态过程）和第二类错误（允许一个失控状态过程继续）都有可能发生。

表 19-1　统计过程控制的结果

		生产过程的状态	
		H_0 为真 过程在于控	H_0 为伪 过程失控
结论	继续过程	正确结论	第二类错误 （允许一个失控状态过程继续）
	调整过程	第一类错误 （调整一个在控状态过程）	正确结论

19.2.1　控制图

控制图（control chart）[⊜]对确定产品中的质量变异是来源于一般原因（在控）还是来源于一般原因（失控），提供了一个做出决策的基础。无论任何时候，一旦检测出失控状态，调整或其他纠正措施将使过程重新回到在控状态。

控制图根据它所包含的数据的类型进行分类。如果根据长度、重量或温度等的变量来测量产品质量，则可以使用 $\bar{x}$ **控制图**（$\bar{x}$ chart）。在这种情况下，根据在产品样本中得到的均值来确定继续或调整生产过程。为了对所有控制图介绍一些一般性的概念，我们先考虑 $\bar{x}$ 控制图的一些特征。

图 19-3 显示了 $\bar{x}$ 控制图的一般结构。控制图的中心线表示在控状态过程的均值，垂直线表示要研究变量的测量尺度。每次从生产过程中抽取一个样本，计算出样本均值 $\bar{x}$，然后将表示 $\bar{x}$ 值的数据点标在控制图上。

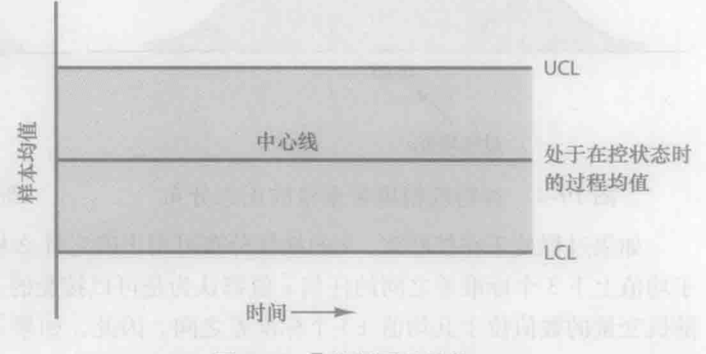


图 19-3　$\bar{x}$ 控制图的结构

记为 UCL 和 LCL 的两条线，对确定过程是处于在控状态还是失控状态非常重要，它们分别被称为控制上限和控制下限。当过程处于在控状态时，$\bar{x}$ 的值位于上下控制限之间的概率很大。位于控制限之外的数值给出强有力的统计证据，过程处于失控状态，应该采取纠正措施。

随着时间的推移，越来越多的数据点被添加到控制图上。数据点的顺序是从左向右，与抽取样本过程的顺序相同。事实上，每当在控制图上标出一个点，就表示我们在进行一次假设检验来确定过程是否处于在控状态。

除 $\bar{x}$ 控制图外，可供使用的其他控制图还有用来监测样本中测量值极差的 ***R* 控制图**（*R* chart），样本中缺陷比例的 ***p* 控制图**（*p* chart）和样本中缺陷项目数量的 ***np* 控制图**（*np* chart）。对每一种情况，控制图都与图 19-3 的 $\bar{x}$ 控制图类似，有一条 LCL，一条中心线，一条 UCL。这些控制图的主要区别是纵轴的测量尺度不同；例如，*p*

⊖ 过程控制程序与本书前面讨论过的假设检验程序密切相关。控制图为处于在控状态过程假设提供了一个持续的检验。
⊜ 根据用连续尺度测量的数据所构造的控制图被称为变量控制图，$\bar{x}$ 控制图就是一个变量控制图。

控制图的测量尺度是样本中有缺陷项目的比率，而不是样本均值。在接下来的讨论中，我们将阐述 $\bar{x}$ 控制图、R 控制图、p 控制图以及 np 控制图的结构和使用。

19.2.2 $\bar{x}$ 控制图：过程的均值和标准差已知

为了阐述 $\bar{x}$ 控制图的结构，让我们考虑 KJW 包装公司的情形。我们记得，KJW 经营一条填装谷物箱的生产线。当生产过程运转正常时，即系统处于在控状态，平均填装重量为 $\mu = 16.05$ 盎司，并且过程的标准差为 $\sigma = 0.10$ 盎司。另外，假设填装重量服从正态分布，如图 19-4 所示。

对于一个处于在控状态的过程，正如第 7 章中介绍的那样，$\bar{x}$ 的抽样分布可以用来确定 $\bar{x}$ 值所期望的变异性。让我们首先简单复习一下 $\bar{x}$ 的抽样分布的性质。首先我们记得，$\bar{x}$ 的期望值或均值等于 μ，μ 是在生产线处于在控状态时的平均填装重量。对容量为 n 的样本，$\bar{x}$ 的标准差（也被称为均值的标准误差），它的计算公式为

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{19-1}$$

另外，因为填装重量服从正态分布，所以对于任何容量的样本，$\bar{x}$ 的抽样分布也服从正态分布。因此，$\bar{x}$ 的抽样分布服从均值为 μ、标准差为 $\sigma_{\bar{x}}$ 的正态分布。如图 19-5 所示。

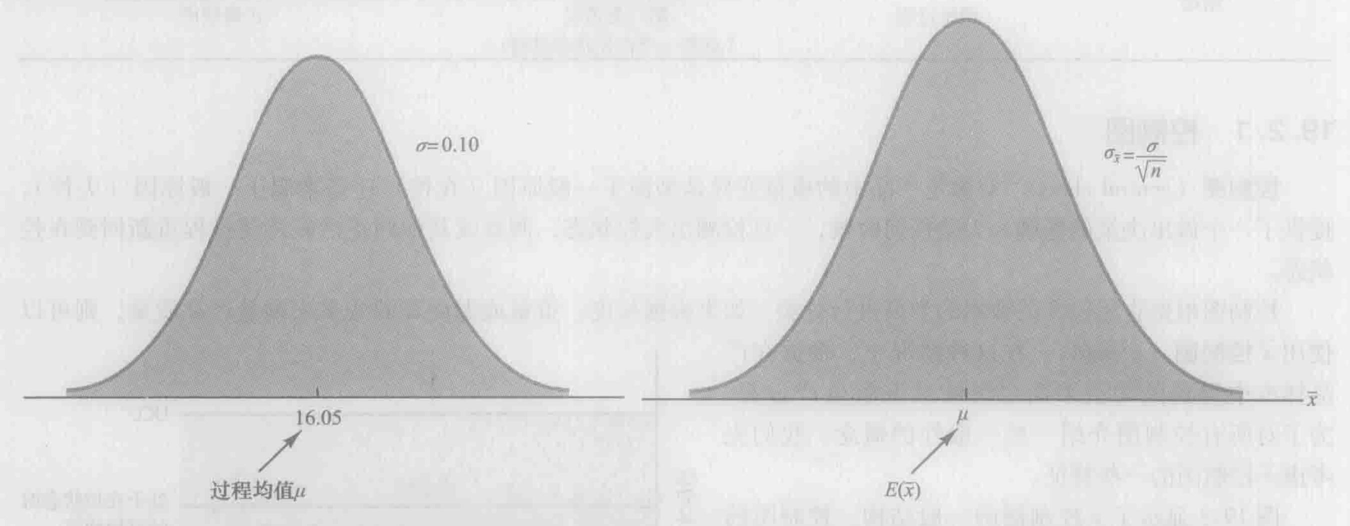


图 19-4　谷物纸箱填装重量的正态分布

图 19-5　对于 n 箱填装的一个样本，重量 $\bar{x}$ 的抽样分布

如果过程处于在控状态，$\bar{x}$ 的抽样分布可用于确定什么样的 $\bar{x}$ 值是可以接受的。质量控制的通常做法是将位于均值上下 3 个标准差之间的任何 $\bar{x}$ 值都认为是可以接受的。回顾正态概率分布的学习，大约 99.7% 的正态分布随机变量的数值位于其均值 ±3 个标准差之间。因此，如果 $\bar{x}$ 的数值位于区间（$\mu - 3\sigma_{\bar{x}}$，$\mu + 3\sigma_{\bar{x}}$）之间，我们将假设过程处于在控状态。因此，$\bar{x}$ 控制图的控制限如下。

$\bar{x}$ 控制图的控制限：过程的均值和标准差已知

$$\text{UCL} = \mu - 3\sigma_{\bar{x}} \tag{19-2}$$

$$\text{LCL} = \mu + 3\sigma_{\bar{x}} \tag{19-3}$$

再次考虑 KJW 包装公司的例子，图 19-4 是填装重量的过程分布，图 19-5 是 $\bar{x}$ 的抽样分布。假设有一个质量控制检查员定期抽取 6 箱作为样本，并且用填装重量的样本均值来确定生产过程是处于在控状态还是失控状态。根据式（19-1），我们得到均值的标准误差为 $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.10/\sqrt{6} = 0.04$。因此，在过程均值等于 16.05 的情况下，控制限为 UCL = 16.05 + 3 × 0.04 = 16.17，LCL = 1605 − 3 × 0.04 = 15.93。图 19-6 是在每 10 小时期间所抽取的 10 个样本结果的控制图。为了便于阅读，将样本标号 1 ~ 10 列在控制图的横轴。

注意图 19-6 中第 5 个样本的均值，有强有力的证据表明，过程处于失控状态。第 5 个样本的均值在 LCL 下

方，存在产品变异性的可指出原因，并出现填装不足的情形。因此，在这个点上应该采取纠正措施，以使过程重新回到在控状态。$\bar{x}$ 控制图上的其他点都位于控制上下限之间的事实，表明纠正措施是成功的。

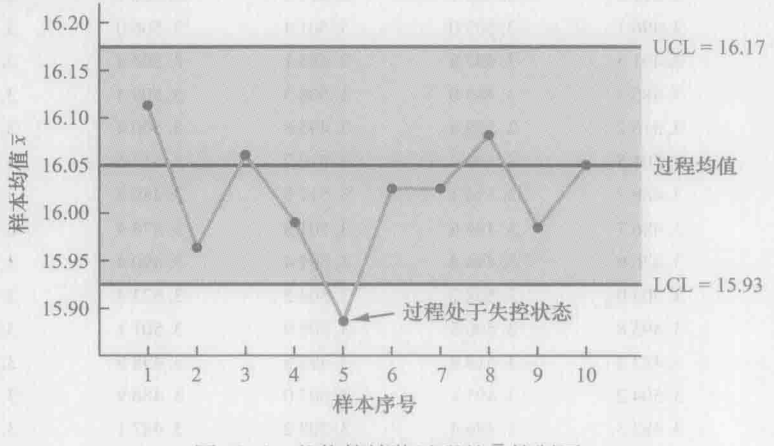


图 19-6　谷物箱填装过程的 $\bar{x}$ 控制图

19.2.3　$\bar{x}$ 控制图：过程的均值和标准差未知

在 KJW 包装公司的例子中，我们介绍了在过程的均值和标准差已知时如何构造 $\bar{x}$ 控制图。在大多数情况下，必须利用样本来估计过程的均值和标准差，而样本是从在控状态的过程中抽取的。例如，KJW 公司在控状态的 10 天里，每天早晨和下午各抽取 5 箱谷物组成一个随机样本。对每一子组或样本，计算他们的样本均值和样本标准差。然后用样本均值和样本标准差的总平均数作为控制图的过程均值和过程标准差。[⊖]

在实践中，最常见的是用极差代替标准差来监测过程的变异性，这是因为极差容易计算。极差可以提供过程标准差很好的估计，因此，极差可以用于构造 $\bar{x}$ 控制图的控制上限和控制下限，而且计算量很少。为了阐明控制图的构造，让我们考虑 Jensen 计算机用品公司所面临的问题。

Jensen 计算机用品供应（JCS）公司生产直径为 3.5 英寸的固态硬盘，并刚刚完成对生产过程的调整使其处于在控状态。假设在生产过程开始的第一个小时抽取 5 张软盘组成一个随机样本，在第二个小时抽取 5 张软盘组成一个随机样本，依此类推，直到抽取了 20 个样本为止。在表 19-2 中给出了每一个样本的软盘直径的样本均值 $\bar{x}_j$ 和极差 R_j。

由总样本均值给出过程均值 μ 的估计值。

总样本均值

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k} \tag{19-4}$$

式中，$\bar{x}_j$ 表示第 j 个样本的均值，$j = 1, 2, \cdots, k$；k 为样本个数。

表 19-2　Jensen 计算机用品公司问题的数据

样本序号	观测值					样本均值 $\bar{x}_j$	样本极差 R_j
1	3.506 5	3.508 6	3.514 4	3.500 9	3.503 0	3.506 5	0.013 5
2	3.488 2	3.508 5	3.488 4	3.525 0	3.503 1	3.502 6	0.036 8
3	3.489 7	3.489 8	3.499 5	3.513 0	3.496 9	3.497 8	0.023 3
4	3.515 3	3.512 0	3.498 9	3.490 0	3.483 7	3.500 0	0.031 6

⊖　对过程的均值和变异性两方面都保持控制是很重要的。

（续）

样本序号	观测值					样本均值 $\bar{x}_j$	样本极差 R_j
5	3.505 9	3.511 3	3.501 1	3.477 3	3.480 1	3.495 1	0.034 0
6	3.497 7	3.496 1	3.505 0	3.501 4	3.506 0	3.501 2	0.009 9
7	3.491 0	3.491 3	3.497 6	3.483 1	3.504 4	3.493 5	0.021 3
8	3.499 1	3.485 3	3.483 0	3.508 3	3.509 4	3.497 0	0.026 4
9	3.509 9	3.516 2	3.522 8	3.495 8	3.500 4	3.509 0	0.027 0
10	3.488 0	3.501 5	3.509 4	3.510 2	3.514 6	3.504 7	0.026 6
11	3.488 1	3.488 7	3.514 1	3.517 5	3.486 3	3.498 9	0.031 2
12	3.504 3	3.486 7	3.494 6	3.501 8	3.478 4	3.493 2	0.025 9
13	3.504 3	3.476 9	3.494 4	3.501 4	3.490 4	3.493 5	0.027 4
14	3.500 4	3.503 0	3.508 2	3.504 5	3.523 4	3.507 9	0.023 0
15	3.484 6	3.493 8	3.506 5	3.508 9	3.501 1	3.499 0	0.024 3
16	3.514 5	3.483 2	3.518 8	3.493 5	3.498 9	3.501 8	0.035 6
17	3.500 4	3.504 2	3.495 4	3.502 0	3.488 9	3.498 2	0.015 3
18	3.495 9	3.482 3	3.496 4	3.508 2	3.487 1	3.494 0	0.025 9
19	3.487 8	3.486 4	3.496 0	3.507 0	3.498 4	3.495 1	0.020 6
20	3.496 9	3.514 4	3.505 3	3.498 5	3.488 5	3.500 7	0.025 9

对表 19-2 的 JCS 数据，总样本均值为 $\bar{\bar{x}} = 3.4995$。这个数值将作为 $\bar{x}$ 控制图的中心线。每一个样本的极差 R_j，是每一个样本的最大值与最小值之差。因此 k 个样本的平均极差计算公式如下。

平均极差

$$\bar{R} = \frac{R_1 + R_2 + \cdots + R_k}{k} \tag{19-5}$$

式中，R_j 表示第 j 个样本的极差，$j = 1, 2, \cdots, k$；k 为样本个数。

对表 19-2 中的 JCS 数据，平均极差为 $\bar{R} = 0.0253$。

前面我们已经表明，$\bar{x}$ 控制图的控制上下限是

$$\bar{\bar{x}} \pm 3\frac{\sigma}{\sqrt{n}} \tag{19-6}$$

因此，为了构造 $\bar{x}$ 控制图的控制限，我们需要估计过程的均值 μ 和标准差 σ，μ 的估计是 $\bar{\bar{x}}$，σ 的估计可以使用极差数据建立。[⊖]

已经证明：过程标准差 σ 的一个估计量为平均极差除以 d_2，d_2 是一个依赖于样本容量 n 的常数。即

$$\sigma \text{ 的估计量} = \frac{\bar{R}}{d_2} \tag{19-7}$$

在表 19-3《美国检验与材料学会数据描述和控制图分析手册》（*American Social for Testing and MaterialsManual on Presentation of Data and Control Chart Analysis*）给出的 d_2 的数值。例如，当 $n = 5$ 时，$d_2 = 2.326$，因此 σ 的估计值为平均极差除以 2.326。如果我们将 $\sigma = \bar{R}/d_2$ 代入式（19-6），则 $\bar{x}$ 控制图的控制限可以写为

$$\bar{\bar{x}} \pm 3\frac{\bar{R}/d_2}{\sqrt{n}} = \bar{\bar{x}} \pm \frac{3}{d_2\sqrt{n}}\bar{R} = \bar{\bar{x}} \pm A_2\bar{R} \tag{19-8}$$

注意，$A_2 = 3/(d_2\sqrt{n})$ 是一个仅依赖于样本容量的常数。表 19-3 中给出 A_2 的数值。对 $n = 5$，$A_2 = 0.577$。因此，$\bar{x}$ 控制图的控制限为

$$3.4995 \pm 0.577 \times 0.0253 = 3.4995 \pm 0.0146$$

⊖ 总样本均值 $\bar{\bar{x}}$ 给出 μ 的估计值，样本极差用来建立 σ 的估计。

因此，UCL = 3. 514，LCL = 3. 485。

表 19-3 $\bar{x}$ 和 R 控制图的因子

样本容量 n	d_2	A_2	d_3	D_3	D_4
2	1. 128	1. 88	0. 853	0	3. 267
3	1. 693	1. 023	0. 888	0	2. 574
4	2. 059	0. 729	0. 880	0	2. 282
5	2. 326	0. 577	0. 864	0	2. 114
6	2. 534	0. 483	0. 848	0	2. 004
7	2. 704	0. 419	0. 833	0. 076	1. 924
8	2. 847	0. 373	0. 820	0. 136	1. 864
9	2. 97	0. 337	0. 808	0. 184	1. 816
10	3. 078	0. 308	0. 797	0. 223	1. 777
11	3. 173	0. 285	0. 787	0. 256	1. 744
12	3. 258	0. 266	0. 778	0. 283	1. 717
13	3. 336	0. 249	0. 770	0. 307	1. 693
14	3. 407	0. 235	0. 763	0. 328	1. 672
15	3. 472	0. 223	0. 756	0. 347	1. 653
16	3. 532	0. 212	0. 750	0. 363	1. 637
17	3. 588	0. 203	0. 744	0. 378	1. 622
18	3. 64	0. 194	0. 739	0. 391	1. 608
19	3. 689	0. 187	0. 734	0. 403	1. 597
20	3. 735	0. 18	0. 729	0. 415	1. 585
21	3. 778	0. 173	0. 724	0. 425	1. 575
22	3. 819	0. 167	0. 720	0. 434	1. 566
23	3. 858	0. 162	0. 716	0. 443	1. 557
24	3. 895	0. 157	0. 712	0. 451	1. 548
25	3. 931	0. 153	0. 708	0. 459	1. 541

资料来源：Adapted from Table 27 of ASTM STP 15D，*ASTM Manual on Presentation of Data and Control Chart Analysis*，Copyright ASTM International，100 Barr Harbor Drive，West Conshohocken，PA19428。

图 19-7 是 Jensen 计算机用品公司问题的 $\bar{x}$ 控制图。我们用表 19-2 中的数据和 StatTools'X/R Chart 程序来构造控制图。[⊖] 中心线为总样本均值$\bar{\bar{x}}$ = 3. 499 5，控制上限（UCL）为 3. 514，控制下限（LCL）为 3. 485。$\bar{x}$ 控制图显示了 20 个样本均值随时间变动的散布情况。由于所有 20 个样本均值都在控制限内，我们确认，抽样期间过程的均值处于在控状态。

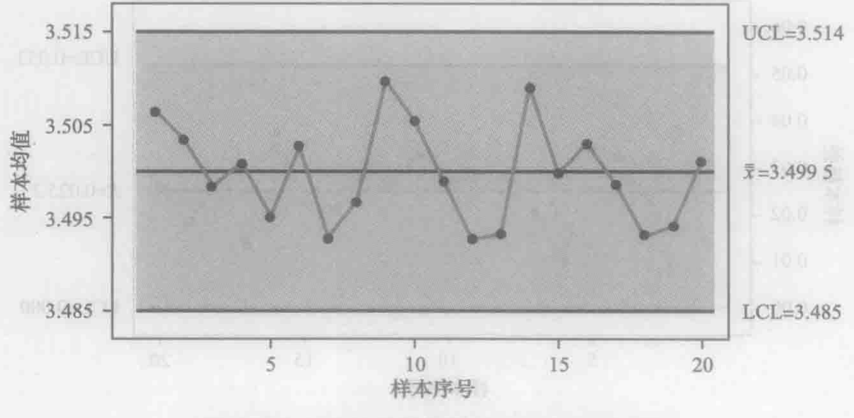


图 19-7 Jensen 计算机用品问题的 $\bar{x}$ 控制图

⊖ 在本章的附录中，我们演示如何利用 StatTools 构造 $\bar{x}$ 和 R 控制图。

19.2.4 *R* 控制图

现在让我们考虑可以用来控制一个过程变异性的极差控制图（*R* 控制图）。为了构造 *R* 控制图，我们需要将样本的极差看作一个有自己的均值和标准差的随机变量。平均极差 $\overline{R}$ 给出了这个随机变量均值的一个估计值。进一步，可以证明，极差标准差的估计是

$$\hat{\sigma}_R = d_3 \frac{\overline{R}}{d_2} \tag{19-9}$$

式中，d_2 和 d_3 是依赖于样本容量的常数，表 19-3 给出了 d_2 和 d_3 的数值。于是，*R* 控制图的 UCL 是

$$\overline{R} + 3\hat{\sigma}_R = \overline{R}\left(1 + 3\frac{d_3}{d_2}\right) \tag{19-10}$$

LCL 是

$$\overline{R} - 3\hat{\sigma}_R = \overline{R}\left(1 - 3\frac{d_3}{d_2}\right) \tag{19-11}$$

如果我们令

$$D_4 = 1 + 3\frac{d_3}{d_2} \tag{19-12}$$

$$D_3 = 1 - 3\frac{d_3}{d_2} \tag{19-13}$$

我们可以将 *R* 控制图的控制限写为

$$\text{UCL} = \overline{R}D_4 \tag{19-14}$$

$$\text{LCL} = \overline{R}D_3 \tag{19-15}$$

表 19-3 也给出了 D_3 和 D_4 的数值。注意，对于 $n = 5$，有 $D_3 = 0$，$D_4 = 2.114$。因此，根据 $\overline{R} = 0.0253$，控制限是

$$\text{UCL} = 0.0253 \times 2.114 = 0.053$$
$$\text{LCL} = 0.0253 \times 0 = 0$$

图 19-8 是 Jensen 计算机用品问题的 *R* 控制图。我们利用表 19-2 中的数据和 StatTools'X/R Chart 程序来构造控制图，控制图的中心线是 20 个样本极差的总平均数 $\overline{R} = 0.0253$。UCL 为 0.053，LCL 为 0.000。*R* 控制图表示了 20 个样本极差随时间变动的散布情况。因为所有 20 个样本的极差都在控制限之内，我们确定，抽样期间过程的变异性处于在控状态[⊖]。

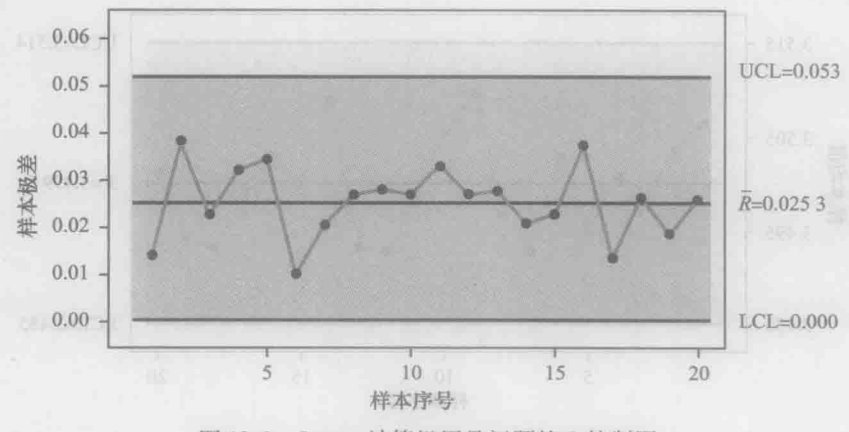


图 19-8 Jensen 计算机用品问题的 *R* 控制图

⊖ 如果 *R* 控制图显示过程处于失控状态，将无法解释 $\bar{x}$ 控制图，直到 *R* 控制图表明过程的变异性处于在控状态。

19.2.5 p 控制图

现在让我们考虑，通过项目是否有缺陷来度量产品质量的情况。以样本中发现的有缺陷项目的比例 $\bar{p}$ 为依据，来确定继续还是调整生产过程。利用有缺陷项目比例的数据构造的控制图称为 p 控制图㊀。

为了说明 p 控制图的结构，让我们考虑邮局使用的自动邮件分拣机的情况。这些自动化机器扫描邮件上的邮政编码，然后将它们分送到合适的邮递线路上去。即使机器运转正常，依然有一些邮件会被分送到错误的邮递线路上去。假设当机器运转正常时，或机器处于在控状态时，有 3% 的邮件被误投。因此，在过程处于在控状态时，错误投递的比例 p 是 0.03。

对于一个处于在控状态的过程，正如第 7 章介绍的那样，$\bar{p}$ 的抽样分布可以用来确定 $\bar{p}$ 值所期望的变异性。我们记得，$\bar{p}$ 的期望值或均值等于 p，即在过程处于在控状态时缺陷项目的比例。对于容量为 n 的样本，$\bar{p}$ 的标准差（又称为比例的标准误差）计算公式是

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{19-16}$$

我们在第 7 章还了解到，当样本容量充分大时，$\bar{p}$ 的抽样分布可以近似服从正态分布。对于 $\bar{p}$，当下面两个条件得到满足时，就可以认为样本容量充分大。

$$np \geqslant 5 \qquad n(1-p) \geqslant 5$$

因此，当样本容量充分大时，$\bar{p}$ 的抽样分布可以近似服从均值为 p，标准差为 $\sigma_{\bar{p}}$ 的正态分布。这个分布如图 19-9 所示。

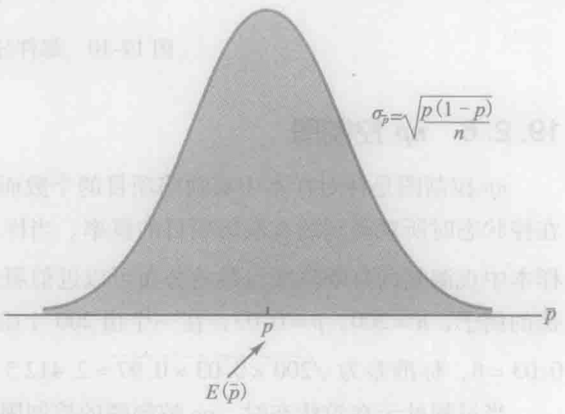


图 19-9 $\bar{p}$ 的抽样分布

为了建立 p 控制图的控制限，我们遵循建立 $\bar{x}$ 控制图的控制限的相同步骤，即在过程处于在控状态时，控制图的控制限设置在缺陷项目比例的上下 3 个标准差或标准误差处。于是，我们得到如下的控制限。

P 控制图的控制限

$$UCL = p + 3\sigma_{\bar{p}} \tag{19-17}$$

$$LCL = p - 3\sigma_{\bar{p}} \tag{19-18}$$

对于 $p=0.03$，样本容量 $n=200$，利用式（19-16），得到标准误差为

$$\sigma_{\bar{p}} = \sqrt{\frac{0.03 \times (1-0.03)}{200}} = 0.0121$$

因此，控制限为 UCL = 0.03 + 3 × 0.012 1 = 0.066 3，LCL = 0.03 − 3 × 0.012 1 = −0.006 3。每当式（19-18）给出 LCL 为负值时，在控制图上令 LCL = 0。

图 19-10 是邮件分拣过程的控制图。所标出的点是从分拣过程抽取的邮件样本中发现的样本缺陷比例。因为所有的点都在控制限之内，因此没有任何证据表明分拣过程处于失控状态。

对于一个处于在控状态的过程，如果缺陷项目的比例未知，则我们首先应该用样本数据来估计它。例如，假设从一个处于在控状态的过程中抽取 k 个容量都为 n 的不同样本。然后确定每个样本缺陷项目的比例。将所有搜集到的数据看作一个大样本，我们计算所有数据的缺陷项目的比例；然后用这个数据估计 p，即在过程处于在控状态时观测到的有缺陷项目的比例。注意，p 的这个估计值也使我们能估计比例的标准误差；从而可以建立控制上、下限。

㊀ 以说明存在缺陷或者缺陷数量为依据的控制图，称为品质控制图。p 控制图是一个品质控制图。

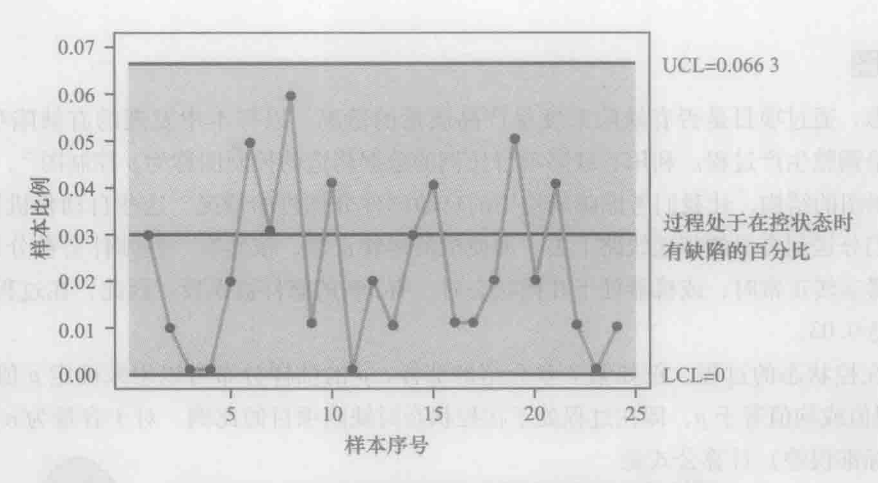


图 19-10 邮件分拣过程中的有缺陷比例的 p 控制图

19.2.6 *np* 控制图

np 控制图是针对样本中有缺陷项目的个数而构造的控制图。在这种情况下，n 为样本容量，p 为当过程处于在控状态时所观测到的有缺陷项目的概率。当样本容量充分大时，即 $np\geqslant5$ 并且 $n(1-p)\geqslant5$，在一个容量为 n 的样本中观测到的有缺陷项目数的分布可以近似服从均值为 np，标准差为 $\sqrt{np(1-p)}$ 的正态分布。因此，对邮件分拣的例子，$n=200$，$p=0.03$，在一个由 200 个邮件组成样本中，观测到有缺陷项目数可以近似服从均值为 $200\times0.03=6$，标准差为 $\sqrt{200\times0.03\times0.97}=2.4125$ 的正态分布。

当过程处于在控状态时，*np* 控制图的控制限为观测到的有缺陷项目的期望数的上下 3 个标准差，于是，我们可以得到如下的控制限。

np 控制图的控制限

$$UCL = np + 3\sqrt{np(1-p)} \tag{19-19}$$

$$LCL = np - 3\sqrt{np(1-p)} \tag{19-20}$$

对于邮件分拣过程的例子，$p=0.03$，$n=200$，控制限是 $UCL=6+3\times2.4125=13.2375$，$LCL=6+3\times2.4125=-1.2375$。当 LCL 为负值时，在控制图上令 $LCL=0$。因此，如果被投送到错误线路的邮件超过 13 件时，我们可以断定，过程处于失控状态。

np 控制图提供的信息与 p 控制图所提供的信息相同。唯一的区别是，*np* 控制图是观测到的有缺陷项目数的分布图，而 p 控制图是观测到的有缺陷项目比例的分布图。因此，如果我们以 p 控制图为依据做出某个特定过程处于失控状态的结论，则我们也能以 *np* 控制图为依据做出该过程处于失控状态的结论。

19.2.7 控制图的解释

根据在控制图中点的位置和分布模式，我们可以确定一个过程是否处于统计控制状态，而且错判的概率很小。过程有可能处于失控状态的主要标志是数据点落在控制限之外，例如图 19-6 中的第 5 个点。发现这样的点是过程处于失控状态的统计证据，在这种情况下，应该尽可能地采取纠正措施。

除了数据点位于控制限之外，某些在控制限之内的数据点的分布模式也可能是质量控制问题的预警信号。例如，假设所有的数据点都在控制限之内，但是大多数的点都位于中心线的一侧。这种分布模式可能说明，或者设备出现了问题，或者原材料有了变化，或者出现了其他质量变化的可指出原因。我们应该对生产过程进行仔细检查以确定质量是否发生了变化。

在控制图看到的另一种模式是随着时间的推移而发生的逐渐变化或趋势。例如，由于工具的磨损，使制造零

件的尺寸逐渐偏离设计的水平；温度或湿度的逐渐变化；设备的整体老化；污垢的堆积或者操作人员的疲劳等，这些也可能使控制图出现一个趋势模式。若连续6~7个数据点呈现单调上升或者单调下降的趋势，就应该引起注意，即使这些数据点都在控制限之内[㊀]。当这种模式出现时，就应该检查生产过程以防备质量上可能的变化或转变，并采取必要的纠正措施。

注释和评论

1. 因为 $\bar{x}$ 控制图的控制限依赖于平均极差的数值，所以这些控制限没有太多的意义，除非过程的变异性处于在控状态。在实践中，通常在构造 $\bar{x}$ 控制图之前构造 R 控制图；如果 R 控制图表明，过程的变异性处于在控状态，然后构造 $\bar{x}$ 控制图。StatTools X/R Chart 选项可以同时给出 $\bar{x}$ 控制图和 R 控制图，本章附录中描述了这一过程的步骤。
2. np 控制图根据有缺陷的数量来监控过程。摩托罗拉的六西格玛质量水平设置的生产目标是：每百万次操作缺陷不超过3.4次；这个目标意味着 $p=0.000\,003\,4$。

练习

方法

2. 从一个处于在控状态的过程中抽取容量为5的25个样本。所有搜集到的数据总和为677.5磅。
   a. 当过程处于在控状态时，过程均值的一个估计值（磅/单位）是多少？
   b. 如果样本容量为5，构造这个过程的 $\bar{x}$ 控制图。假设当该过程处于在控状态时，过程的标准差为0.05，过程的均值为（a）中的估计值。

☆4. 从一个过程中抽取20个样本，每个样本的容量都为8，结果为 $\bar{\bar{x}}=28.5$，$\bar{R}=1.6$。计算这个过程的 $\bar{x}$ 控制图和 R 控制图的上下控制限。

应用

6. 一个质量控制过程检测每箱洗涤剂的重量。控制限设为 UCL = 20.12 盎司，LCL = 19.90 盎司。对抽样和检测过程，样本容量为5。这个制造过程的过程均值和过程标准差为多少？

8. 在生产过程正常运转或处于在控状态的几个星期中，为检测断裂强度抽取了20个样本，每个样本有150包合成纤维网球线。在检测的3 000包中，有141包未能符合制造商规格。
   a. 当系统处于在控状态时，过程的有缺陷比例的估计值是多少？
   b. 计算 p 控制图的上下控制限。
   c. 如果检验了一个150包网球线的新样本，其中12包有缺陷，根据（b）的结果，应该做出什么结论？在这种情况下，是否出现可指出原因？
   d. 计算 np 控制图的上下控制限。
   e. 利用（d）中的结果回答（c）中的问题。
   f. 在这种情况下，哪一个将是首选的控制图？请解释。

19.3 接受抽样

在接受抽样[㊁]中，我们感兴趣的项目可能是正在进货的原材料或外购的零部件，以及来自总装线的制成品。假设我们希望以指定产品的质量特性为依据，决定是否接受或拒绝一组产品项目。在质量控制术语中，一组项目称为**一批**（lot），**接受抽样**（acceptance sampling）是一种统计方法，该方法能使我们将接受 - 拒绝的决定建立在检测从一批中抽取项目样本的基础上。

图19-11是接受抽样的一般步骤。在收到一批产品之后，从中抽取一个项目样本进行检测。将抽测的结果与指定的质量特性进行比较。如果质量特性得到满足，则接受这批产品，并送往生产线或交付给客户。如果拒绝这

㊀ 即使所有的数据点都在上下控制限之内，过程也有可能出于失控状态。样本数据点的趋势或在中心线上下异常长期运动，也可能显示过程处于失控状态。

㊁ 接受抽样比百分之百检测有以下优点：①通常费用较少；②由于较少地搬运和检验，因而产品损坏较少；③需要较少的检验人员；④如果必须要使用破坏性检验，接受抽样是唯一可行的方法。

批产品，管理人员必须作出如何处置这批产品的决定。在一些情况下，可能决定保留这批产品，但是应该剔除无法接受的或者不符合规定的项目；在另一些情况下，可能将这批产品退回给供应商，且所有费用均由供应商支付，额外的工作量和费用记在供应商名下，将激励供应商提供高质量的产品。最后，如果拒绝批次中有制成品，则这些制成品必须报废或者返工，以达到可接受的质量标准。



图 19-11　接受抽样程序

接受抽样统计程序的依据是第 9 章介绍的假设检验方法。原假设与备择假设的陈述如下。

H_0：高质量批

H_a：低质量批

表 19-4 是假设检验程序的结果。注意，正确的决定对应接受一个高质量批，或者拒绝一个低质量批。但是，同其他假设检验程序一样，我们需要注意犯第一类错误（拒绝一个高质量批）或第二类错误（接受一个低质量批）的可能性。

表 19-4　接受抽样的结果

		批的状态	
		H_0 为真 高质量批	H_0 为伪 低质量批
决定	接受该批	正确结论	第二类错误 （接受一个低质量批）
	拒绝该批	第一类错误 （拒绝一个高质量批）	正确结论

第一类错误的概率对批产品的生产者造成了风险，因此被称为**生产者风险**（producer's risk）。例如，生产者风险为 0.05，这意味着被错误地拒绝一个高质量批的可能性为 5%。另一方面，第二类错误的概率对批产品的消费者造成了风险，因此被称为**消费者风险**（consumer's risk）。例如，消费者风险为 0.10，这意味着错误地接受一个低质量批的可能性为 10%，并且这个低质量批被用于生产或卖给消费者。接受抽样程序的设计者可以控制生产者风险和消费者风险的规定值。为了说明如何确定风险值，我们考虑 KALI 有限公司所面临的问题。

19.3.1　KALI 有限公司：接受抽样的实例

KALI 有限公司是一家生产家电的公司，该公司在市场上销售各种名称的产品。但是，KALI 并不能制造其产品所用的每一个部件，因此，需要直接从供应商处购买一些部件。例如，KALI 购买的一个部件是用在家庭空调器上的超载保护器，它是一种保护装置，能在空调器过热时关闭压缩机。如果超载保护器运转不正常，则空调压缩机有可能受到严重损坏。因此，KALI 很关注超载保护器的质量。保证质量的一种方法是检测所收到的每一个部件，这种方法被称为 100% 检验法。但是，为了确定一个超载保护器正常运行，必须对部件进行长时间的、费用昂贵的检验，而且 KALI 不可能检测收到的每一个超载保护器。

作为替代方法，KALI 利用接受抽样方案来检测超载保护器的质量。接受抽样方案要求 KALI 的质量控制检查员从每批产品中抽取样本并进行检验。如果在样本中发现极少有缺陷的产品，则该批可能是高质量，应该接收。但是，如果在样本中发现大量有缺陷的产品，则该批可能是低质量，应该拒收。

一个接受抽样方案包含样本容量 n 和接收准则 c。**接收准则**（acceptance criterion）是在样本中发现有缺陷项目的最大数量，并且仍然可以接收该批。例如，对 KALI 公司的问题，我们假设从装运的产品或批中每次抽取 15 个项目组成样本。此外，假设质量控制人员表示，仅在没有发现有缺陷项目的情况下才接收该批。在这种情况下，由质量控制人员建立的接受抽样方案为 $n=15$，$c=0$。

这个接受抽样方案对质量控制检查员来说是很容易实施的。质检员只需要抽取 15 个项目组成样本，并对每

一个项目进行检验，并根据下面的决策规则作出结论。

- 如果没有发现有缺陷项目，则接收该批。
- 如果发现一个或一个以上有缺陷项目，则拒收该批。

在实施接受抽样方案之前，质量控制人员希望评估该方案的风险或错误概率。只有生产者风险（第一类错误）和消费者风险（第二类错误）都控制在一个合理的水平时，这个方案才能得以实施。

19.3.2 计算接收一批的概率

分析生产者风险和消费者风险的关键是“如果……，则……”类型的分析，即我们将假设，已知一批中有缺陷项目的百分比，对于给定的抽样方案，计算接收这批的概率。通过改变假设有缺陷项目的百分比，我们可以检验抽样方案关于两类风险的影响。

假设我们已经收到大批量的超载保护器，且其中 5% 的超载保护器有缺陷。对有 5% 缺陷项目的装运产品或批，$n=15$，$c=0$ 的抽样方案将使我们接收这批的概率是多大？因为每个被检测的超载保护器，要么有缺陷，要么没有缺陷，而且批的容量很大。因此，对于容量为 15 的样本，有缺陷项目数服从一个二项分布。二项概率函数在第 5 章中介绍过，其形式如下。

接受抽样的二项概率函数

$$f(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{(n-x)} \tag{19-21}$$

式中，n 为样本容量；p 为批中有缺陷项目的比例；x 为样本中有缺陷项目的数量；$f(x)$ 为样本中有 x 个有缺陷项目的概率。

对 KALL 的接受抽样方案，$n=15$。于是，对有缺陷项目为 5%（$p=0.05$）的一批，我们有

$$f(x) = \frac{15!}{x!(15-x)!}0.05^x(1-0.05)^{(15-x)} \tag{19-22}$$

利用式（19-22），$f(0)$ 为 0 个超载保护器有缺陷，并且接收该批的概率。在式（19-22）中，由于 $0!=1$，因此有

$$f(0) = \frac{15!}{0!15!}0.05^0(1-0.05)^{(15-0)} = \frac{15!}{0!15!}0.05^0 0.95^{15} = 0.95^{15} = 0.4633$$

我们现在知道，对 $n=15$，$c=0$ 的抽样方案，有缺陷项目为 5% 的一批，接收的概率为 0.463 3。因此，有缺陷项目为 5% 的一批，拒收的概率相应的为 $1-0.4633=0.5367$。

可以使用 Excel 的 BINOM. DIST 函数，使这些二项概率的计算简单化。利用这个函数，我们能确定，如果有缺陷项目为 10%，对于 $n=15$，$c=0$ 的抽样方案，接收这批的概率为 0.205 9。如果有缺陷项目分别为 1%，2%，3%，…，对于 $n=15$，$c=0$ 的抽样方案，接收这批的概率汇总如表 19-5 所示。

表 19-5　对 KALI 问题 $n=15$，$c=0$ 的接收一批的概率

批中有缺陷的百分比	接收一批的概率
1	0.860 1
2	0.738 6
3	0.633 3
4	0.542 1
5	0.463 3
10	0.205 9
15	0.087 4
20	0.035 2
25	0.013 4

根据表 19-5 的概率，将接收一批的概率和批中有缺陷项目的百分比绘成图 19-12。这个图形或曲线，被称为 $n=15$，$c=0$ 接受抽样方案的**抽样特性曲线**（operating characteristic，OC）。

或许我们应该考虑其他抽样方案，它们有不同的样本容量 n 或不同的接收准则 c。首先考虑样本容量仍保持 15 不变，但是接收准则从 $c=0$ 增加为 $c=1$ 的情形，即如果在样本中发现 0 个或 1 个有缺陷部件，我们将接收该批。对于有缺陷项目为 5%（$p=0.05$）的一批，可以利用式（19-21）的二项概率函数，或 Excel 的 BINOM. DIST 函数，计算 $f(0)=0.4633$，$f(1)=0.3658$。因此，对 $n=15$，$c=1$ 抽样方案，接收有缺陷项目为 5% 的一批的概率为 0.463 3 +

0.365 8 = 0.829 1。

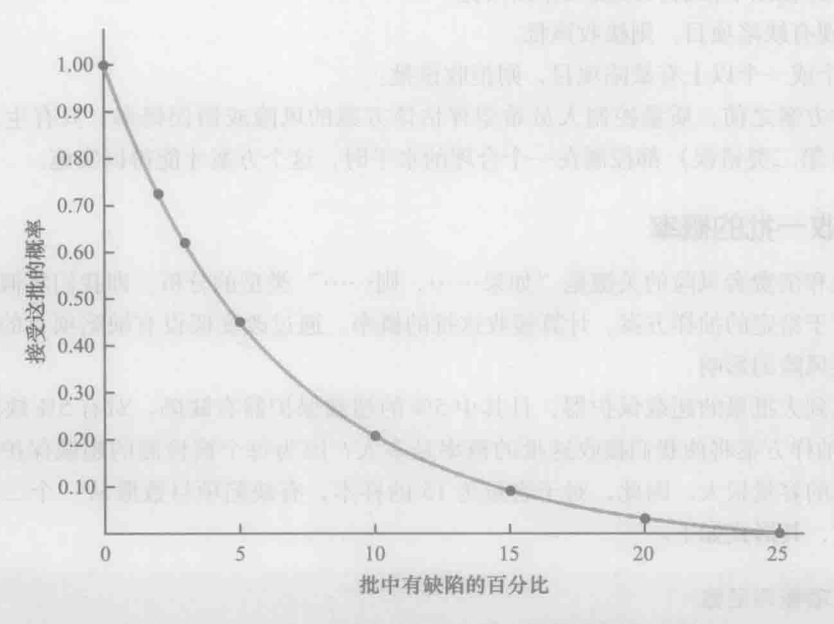


图 19-12 $n=15$，$c=0$ 的接受抽样方案的抽样特性曲线

继续这样的计算，我们得到图 19-13。该图是 KALI 公司问题的四种接受抽样方案的抽样特性曲线。分别考虑容量为 15 和 20 的样本，我们注意到，无论批中有缺陷项目的比例如何，$n=15$，$c=1$ 的抽样方案给出接收该批的概率最大。$n=20$，$c=0$ 的抽样方案给出接收该批的概率最小，但该方案也给出了拒绝该批的概率最大。

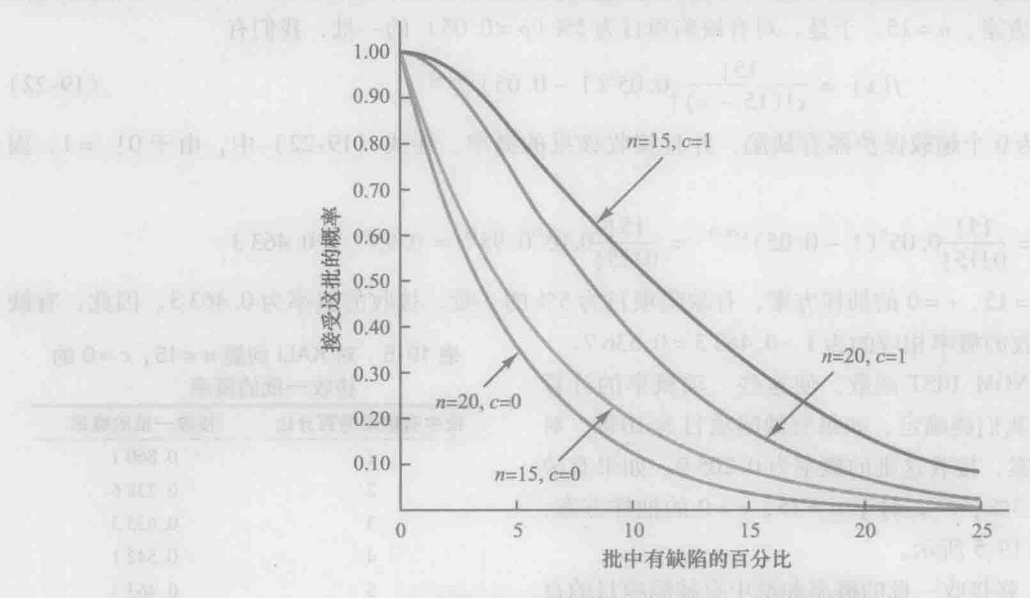


图 19-13 四种接受抽样方案的抽样特性曲线

19.3.3 选择接受抽样方案

现在我们已经知道如何使用二项分布，根据一个给定的有缺陷项目的比例来计算接收一批的概率。对于正在研究的应用，我们已经选好 n 和 c 的数值来确定所需要的接受抽样方案。为了制定这个方案，管理人员必须对批中有缺陷项目的比例指定两个数值。一个数值 p_0 用于控制生产者风险；另一个数值 p_1 用于控制消费者风险。

我们使用如下记号：

α—— 生产者风险；拒收有缺陷比例为 p_0 的一批的概率

β—— 消费者风险；接收有缺陷比例为 p_1 的一批的概率

假设对 KALI 公司的问题，管理人员规定 $p_0 = 0.03$，$p_1 = 0.15$。从图 19-14 的 $n = 15$，$c = 0$ 的抽样特征曲线中，我们看到 $p_0 = 0.03$ 给出的生产者风险大约为 $1 - 0.63 = 0.37$，$p_1 = 0.15$ 给出的消费者风险大约似为 0.09。因此，如果管理人员愿意承受拒收有缺陷项目比例为 3% 的一批的概率为 0.37（生产者风险），同时承受接收有缺陷项目比例为 15% 的一批的概率为 0.09（消费者风险），则 $n = 15$，$c = 0$ 的接受抽样方案是可以接受的。

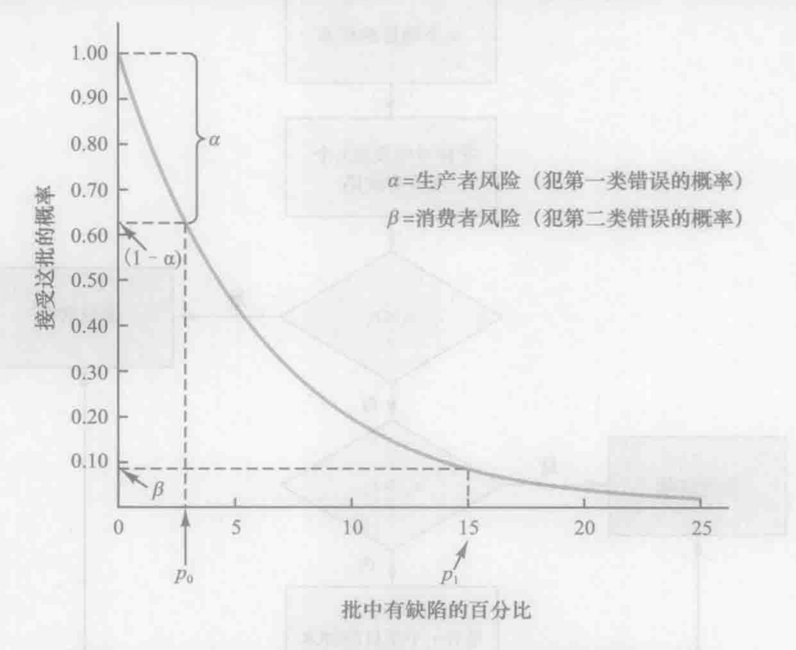


图 19-14　$n = 15$，$c = 0$，$p_0 = 0.03$，$p_1 = 0.15$ 的抽样特性曲线

但是，假如管理人员要求生产者风险 $\alpha = 0.10$，消费者风险 $\beta = 0.19$。我们看到，$n = 15$，$c = 0$ 的抽样方案有一个比消费者预期的更小的风险，但这也有一个不可接受的大的生产者风险。$\alpha = 0.37$ 的事实表明，当一批只有 3% 的项目有缺陷时，错误地拒收该批的概率为 37%。由此可见生产者风险过高，我们应该考虑其他的接受抽样方案。

利用 $p_0 = 0.03$，$\alpha = 0.10$，$p_1 = 0.15$ 和 $\beta = 0.20$，图 19-13 表明 $n = 20$，$c = 1$ 的接受抽样方案最接近于同时满足生产者风险和消费者风险的需要[⊖]。

正如本节所介绍的，在决定抽样方案时，我们需要结合预期的生产者风险和消费者风险，考虑一些计算或抽样特性曲线。幸运的是，我们已经出版了一些抽样方案表。例如，美国军用标准表（MIL-STD-105D），在设计接受抽样方案时提供了很多有用的信息。列在本书参考文献上的关于质量控制的较高级的教材介绍了这些表的使用方法。较高级的教材也讨论了抽样成本在确定最优抽样方案中的作用。

19.3.4　多重抽样方案

我们介绍 KALI 公司问题的接受抽样程序是单样本方案。它之所以被称为单样本方案，是因为仅仅用了一个样本或一个抽样阶段。在确定了样本中有缺陷部件的个数后，就必须做出接收或拒收该批的结论。另一种可供选择的抽样方案是**多重抽样方案**（multiple sampling plan），它涉及两个或多个抽样阶段。在每一个阶段有 3 种可能的结论：停止抽样并接收该批、停止抽样并拒收该批，或者继续抽样。尽管这些做法比较复杂，但是对于同样的 α 和 β 的概率，多重抽样方案的总样本容量往往比单样本抽样方案要小。

⊖　本节后的练习第 13 题将要求你对 $n = 20$，$c = 1$ 的接受抽样方案，计算生产者风险和消费者风险。

图 19-15 是两阶段或双重抽样方案的流程。首先选择 n_1 个项目组成第一个样本，如果有缺陷部件的数量 $x_1 \leqslant c_1$，则接收该批；如果 $x_1 \geqslant c_2$，则拒收该批；如果 x_1 介于 c_1 和 c_2 之间（即 $c_1 < x_1 < c_2$），则选择 n_2 个项目组成第二个样本。由第一个样本有缺陷部件的数量 x_1 和第二个样本有缺陷部件的数量 x_2，可以确定有缺陷部件的总数 $x_1 + x_2$。如果 $x_1 + x_2 \leqslant c_3$，则接收该批；否则拒收该批。双重抽样方案的建立更加困难，因为样本容量 n_1 和 n_2 及接收准则 c_1，c_2 和 c_3 必须同时满足生产者和消费者所预期的双重风险。

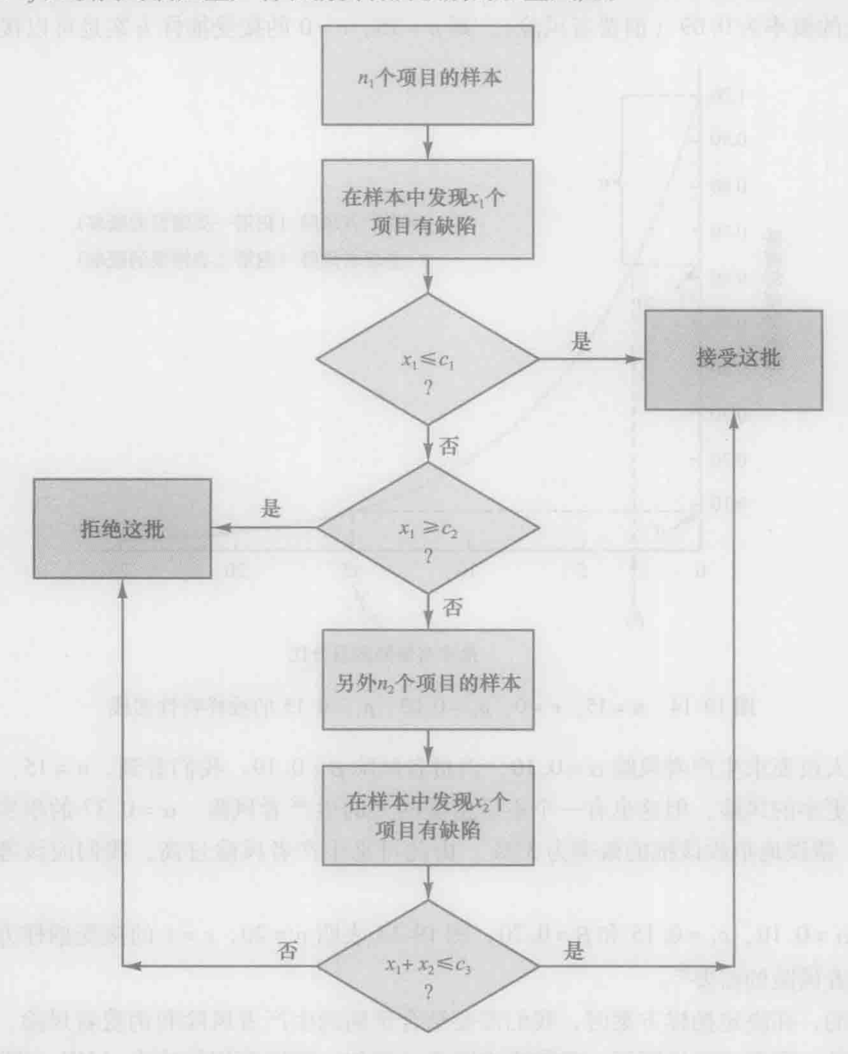


图 19-15　两阶段接受抽样方案

注释和评论

1. 对于接受抽样，使用二项分布的依据是假设：批的容量较大。如果批容量较小，超几何分布是合适的分布。
2. 在 MIL-ST-105D 的抽样表中，p_0 被称为可接受质量水平（AQL）。在一些抽样表中，p_1 被称为批容许缺陷百分数（LTPD）或拒绝质量水平（RQL）。许多出版的抽样方案也使用了一些质量指数，如无差异质量水平（IQL）和平均出厂质量界限（AOQL）。在本书参考文献中列出的更为高级的教材给出了这些质量指数的全面讨论。
3. 在本节中我们介绍了品质抽样方案。在这些方案中，被抽中的每一个项目分为有缺陷和无缺陷。在变量抽样方案中，我们抽取一个样本并测量它的质量特性。例如，对黄金珠宝质量的检测可以是它的含金量。计算样本统计量，例如样本中黄金的平均含量，并将它与一个允许值进行比较，来确定是接受还是拒收该批产品。

练　习

方法

☆10. 对 $n=25$，$c=0$ 的接受抽样方案，计算接收缺陷率为 2% 的一批的概率。如果缺陷率为 6%，接收该批的概率为多少？

12. 对 $n=20$，$c=1$ 的接受抽样方案，重复第 11 题的计算。当接收准则 c 的数值增加时，生产者风险会出现什么变化？请解释。

应用

14. 为了检验刚刚装运的原材料，制造商正在考虑容量为 10，15 和 20 的样本。利用附录 B 中表 5 的二项概率，选择一个抽样方案，要求满足：当 $p_0=0.05$ 时，生产者风险为 $\alpha=0.03$；当 $p_1=0.30$ 时，消费者风险为 $\beta=0.12$。

小　结

本章我们讨论了统计方法如何用来帮助进行质量控制。首先我们介绍了作为图示法的 $\bar{x}$，R，p 和 np 控制图来检测生产过程。对每种控制图确定各自的控制限；定期抽取样本，并将数据点描绘在控制图上。当数据点位于控制限之外时，则表明过程处于失控状态，必须采取纠正措施。在控制限之内的数据点的模式，也能指出潜在的质量控制问题，纠正措施的建议可能是必要的。

我们还讨论了被称为接受抽样的技术，根据接受抽样程序抽取并检验一个样本。样本中有缺陷项目的个数提供了接收或拒收一批的根据。为了控制生产者风险（第一类错误）和消费者风险（第二类错误）可以调整样本容量和接收准则。

关键术语

total quality (TQ)　**全面质量**　通过不断改进和学习的策略以提高消费者满意度和降低实际成本的一种全面的系统方法。

six sigma　**六西格玛**　用测量和统计分析得到一个高质量水平的方法，它使得每百万出现缺陷的机会不超过 3.4。

quality control　**质量控制**　确定是否达到质量标准的一系列检查和测量行为。

assignable Causes　**可指出的原因**　由于诸如工具的磨损、不正确的机器安装、低质量的原材料或者操作人员的失误等因素，使得过程的产品出现变异性。一旦发现产品变异性的可指出的原因，就应采取纠正措施。

common Causes　**一般原因**　由于随机性，使得过程的产品出现正常的或自然的变异性。当产品的变异性来源于一般原因时，不需要采取纠正措施。

control chart　**控制图**　用于帮助确定一个过程是处于在控状态还是失控状态的一种图形工具。

$\bar{x}$ chart　**$\bar{x}$ 控制图**　根据一个变量，如长度、重量、温度等的均值来检测一个过程的产品质量时所使用的控制图。

R chart　**R 控制图**　根据一个变量的极差来检测一个过程的产品质量时所使用的控制图。

p chart　**p 控制图**　根据缺陷率来检测一个过程的产品质量时所使用的控制图。

np chart　**np 控制图**　根据有缺陷项目的数量来检测一个过程的产品质量时所使用的控制图。

lot　**批**　一组项目，比如正在进货的原材料或外购的零部件以及来自总装线的制成品。

acceptance sampling　**接受抽样**　通过在样本中发现有缺陷项目的数量来决定是接收还是拒收一批的统计方法。

producer's risk　**生产者风险**　拒绝一个高质量批的风险，也称为第一类错误。

consumer's risk　**消费者风险**　接受一个低质量批的风险，也称为第二类错误。

acceptance criterion　**接收准则**　在样本中发现有缺陷项目的最大数量，并且仍然可以接收该批。

operating characteristic curve　**抽样特性曲线**　显示接受一批概率为批中有缺陷项目百分比的函数的图形。这条曲线可以用于确定，某个特定的接受抽样方案是否同时满足生产者风险和消费者风险的要求。

multiple sampling plan　**多重抽样方案**　利用不止一个样本或一个抽样阶段的接受抽样方案。根据在样本中发现的有缺陷项目的数量，来决定是接收该批、拒收该批，还是继续抽样。

重要公式

均值的标准误差

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (19\text{-}1)$$

$\bar{x}$ 控制图的控制限：过程的均值和标准差已知

$$UCL = \mu + 3\sigma_{\bar{x}} \quad (19\text{-}2)$$

$$LCL = \mu - 3\sigma_{\bar{x}} \quad (19\text{-}3)$$

总样本均值

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k} \quad (19\text{-}4)$$

平均极差

$$\bar{R} = \frac{R_1 + R_2 + \cdots + R_k}{k} \quad (19\text{-}5)$$

$\bar{x}$ 控制图的控制限：过程的均值和标准差未知

$$\bar{\bar{x}} \pm A_2\bar{R} \quad (19\text{-}8)$$

R 控制图的控制限

$$UCL = \bar{R}D_4 \quad (19\text{-}14)$$

$$LCL = \bar{R}D_3 \quad (19\text{-}15)$$

比例的标准误差

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \quad (19\text{-}16)$$

p 控制图的控制限

$$UCL = p + 3\sigma_{\bar{p}} \quad (19\text{-}17)$$

$$LCL = p - 3\sigma_{\bar{p}} \quad (19\text{-}18)$$

np 控制图的控制限

$$UCL = np + 3\sqrt{np(1-p)} \quad (19\text{-}19)$$

$$LCL = np - 3\sqrt{np(1-p)} \quad (19\text{-}20)$$

接受抽样的二项概率函数

$$f(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{(n-x)} \quad (19\text{-}21)$$

补充练习

16. 从一个处于在控状态的生产过程抽取20个容量为5的样本，得到如下的样本均值资料。

95.72 95.24 95.18 95.44 95.46 95.32
95.40 95.44 95.08 95.50
95.80 95.22 95.56 95.22 95.04 95.72
94.82 95.46 95.60 95.78

a. 根据这些数据，当过程处于在控状态时，均值的估计值是多少？

b. 假设过程的标准差 $\sigma=0.50$，构造该生产过程的控制图，假设过程的均值为（a）中的估计值。

c. 这20个样本的均值是否表明过程处于失控状态？

18. 抽取25个容量为5的样本，得到 $\bar{\bar{x}}=5.42$，$\bar{R}=2.0$。计算 $\bar{x}$ 和 R 控制图的控制限，并估计过程的标准差。

20. 下面是搜集到的 Master Blend Coffee 的生产过程的数据。数据显示的是3磅重的听装咖啡的填装重量。利用这些数据构造 $\bar{x}$ 和 R 控制图。对生产过程的质量可以得出什么结论？

样本序号	观测值				
	1	2	3	4	5
1	3.05	3.08	3.07	3.11	3.11
2	3.13	3.07	3.05	3.10	3.10
3	3.06	3.04	3.12	3.11	3.10
4	3.09	3.08	3.09	3.09	3.07
5	3.10	3.06	3.06	3.07	3.08
6	3.08	3.10	3.13	3.03	3.06
7	3.06	3.06	3.08	3.10	3.08
8	3.11	3.08	3.07	3.07	3.07
9	3.09	3.09	3.08	3.07	3.09
10	3.06	3.11	3.07	3.09	3.07

22. 1 200个零售商店的管理人员从中心供应仓库每月做两次进货订单。过去的经验表明，4%的订单至少有1个错误，如产品装运错误、装运数量错误和漏装了已经预订的产品。每个月抽取200份订单组成随机样本，并进行准确性检验。

a. 构造这种情况控制图。

b. 下面是6个月中至少有1个错误的订单数：10，15，6，13，8和17。将这些数据绘在控制图上。对于订单过程，你的控制图说明什么问题？

24. 设计的接受抽样方案为 $n=15$，$c=1$，生产者风险为0.075。

a. p_0 的值是0.01，0.02，0.03，0.04或0.05吗？这个值意味着什么？

b. 如果 $p_1=0.25$，则这个方案的消费者风险是多少？

第20章 指数

CHAPTER 20

实践中的统计

美国劳工部，劳工统计局

华盛顿特区

美国劳工部通过其下属的劳工统计局编纂并发布了各类指数及其他统计资料，以此作为反映美国商业和经济活动的晴雨表。例如，劳工统计局编纂和出版消费者价格指数（CPI）、生产者价格指数（PPI）以及各类工人的平均工作时间与收入等资料。在劳工统计局编制的指数中，应用最广泛的是消费者价格指数，它常常被用于测量通货膨胀。

2009 年 3 月美国劳工统计局的报告显示，2 月的消费者价格指数上涨了 0.5%。2 月消费者价格指数的水平为 212.2，与 2008 年同期相比高出了 0.3%。经过季节调整后，消费者价格指数在 1 月上涨 0.3% 后，2 月又上涨了 0.4%。汽油价格指数增长 8.3% 似乎是上涨的主要因素，而食品类指数实际上下降 0.1%。有些经济学家认为消费者价格指数上涨是好消息，因为它可能缩短了通货紧缩的周期。

几天前，劳工统计局的报告显示，经过季节调整后的生产者价格指数 2 月上涨了 0.1%，这是在 1 月上涨 0.8% 和前一年 12 月下降 1.9% 后的上涨。生产者价格指数用于测量批发市场的价格变动，也常常被看作是消费者价格指数变动的导向指标。2 月较慢的增长率受到能源商品增长率逐渐减少的严重影响。能源商品指数在 1 月上涨 3.7% 后，2 月上涨了 1.3%。

本章我们将介绍各种指数的计算，例如消费者价格指数、生产者价格指数等，并对它们加以说明。

美国政府每个月都发布各类指数，编制这些指数是为了帮助居民了解当前的商业和经济状况。这些指数中被广为了解和运用的是消费者价格指数（CPI）。顾名思义，消费者价格指数是消费者为购买的物品支付的价格变动的指示器。特别地，消费者价格指数可以度量一段时间内的价格变动。给定一个开始时期，也称为基期，其指数为 100，消费者价格指数就能对报告期的消费价格与基期的消费价格进行比较。例如，消费者价格指数为 125，意味着报告期的消费价格整体上比基期大约提高 25%。尽管只有少数居民能确切理解这个数字的含义，但他们的确知道这个指数，知道该数值的增加意味着价格的上涨。

虽然消费者价格指数可能是最为著名的指数，但是还有许多其他政府和私人机构的指数，它们也可以帮助我们度量和理解如何将一个时期的经济状况和另一个时期的经济状况进行比较。本章的目的就在于阐述广为使用的几种指数。我们通过编制一些简单的指数，使我们对如何计算指数有一个更好的理解。

20.1 价比

价格指数最简单的形式，就是一个给定商品项目的报告期单价与基期单价的简单对比。例如，表 20-1 是 1990 ~ 2011 年每加仑普通汽油的成本资料，为了方便与其他年份的比较，每加仑汽油的现实成本数据可转化为一个**价比**（price relative），即将每一时期的单价表示为基期单价一个百分数。

表 20-1 普通汽油的成本（全配方）

年份	每加仑的价格（美元）	年份	每加仑的价格（美元）	年份	每加仑的价格（美元）	年份	每加仑的价格（美元）
1990	1.30	1996	1.22	2002	1.34	2008	3.25
1991	1.10	1997	1.20	2003	1.56	2009	2.35
1992	1.09	1998	1.03	2004	1.85	2010	2.78
1993	1.07	1999	1.14	2005	2.27	2011	3.52
1994	1.08	2000	1.48	2006	2.57		
1995	1.11	2001	1.42	2007	2.80		

资料来源：U.S. Energy Information Administration.

$$t\text{期的价比} = \frac{t\text{期的价格}}{\text{基期的价格}} \times 100 \tag{20-1}$$

根据表20-1的资料，以1990年为基期，可以算出从1990~2011年每加仑普通汽油的价比，将它们列在表20-2中。通过了解价比，任何时期的价格就很容易与基期的价格进行比较。例如，1995年的价比为85.4，表明汽油价格1995年比1990年下降14.6%；类似地，2002年的价比为103.1，表明汽油价格2002年比1990年上涨3.1%；2011的价比为280.8，表明汽油价格从基期1990年开始上涨了170.8%。就像普通汽油等现象一样，价比在了解和解释经济与商业状况随时间的变动方面是很有帮助的。

表20-2 一加仑普通汽油的价比

年份	价比（以1990年为基期）	年份	价比（以1990年为基期）
1990	(1.30/1.30)×100=100.0	2001	(1.42/1.30)×100=109.2
1991	(1.10/1.30)×100=84.6	2002	(1.34/1.30)×100=103.1
1992	(1.09/1.30)×100=83.8	2003	(1.56/1.30)×100=120.0
1993	(1.07/1.30)×100=82.3	2004	(1.85/1.30)×100=142.3
1994	(1.08/1.30)×100=83.1	2005	(2.27/1.30)×100=174.6
1995	(1.11/1.30)×100=85.4	2006	(2.57/1.30)×100=197.7
1996	(1.22/1.30)×100=93.8	2007	(2.80/1.30)×100=215.4
1997	(1.20/1.30)×100=92.3	2008	(3.25/1.30)×100=250.0
1998	(1.03/1.30)×100=79.2	2009	(2.35/1.30)×100=180.8
1999	(1.14/1.30)×100=87.7	2010	(2.78/1.30)×100=213.8
2000	(1.48/1.30)×100=113.8	2011	(3.52/1.30)×100=270.8

20.2 综合物价指数

尽管价比能用来解释单个商品项目的价格随时间的变动情况，但我们通常更感兴趣的是一组商品项目整体上的价格变动。例如，我们想要一个指数来全面测量生活费用随时间变动的情况。我们希望这个指数以各种生活用品的价格变动为依据，包括食品、住房、服装、交通运输和医疗保健等。**综合价格指数**（aggregate price index）就是为了度量一组商品项目的综合变动这一特殊目的而编制的。

以普通汽车运行费用的一组分类项目为例，我们考虑综合物价指数的编制。为了方便起见，我们只考虑汽油、润滑油、轮胎和保险费用四个项目。

表20-3是1990~2011年上述四个项目的汽车运行费用的资料。以1990年为基期的四个项目的综合物价指数，将给我们一个1990~2011年普通汽车运行费用变动情况的测度。

表20-3 汽车运行费用指数的数据

商品项目	单价（美元）	
	1990年	2011年
汽油（加仑）	1.30	3.52
润滑油（夸脱）	2.10	6.25
轮胎	130.00	145.00
保险费用	820.00	1 040.00

未加权综合物价指数是报告期（如2011年）单价之和与基期（如1990年）单价之和的对比。令

P_{it}表示t期第i种商品的单价，P_{i0}表示基期第i种商品的单价。则未加权t期的综合物价指数I_t可表示为

$$I_t = \frac{\sum P_{it}}{\sum P_{i0}} \times 100 \tag{20-2}$$

式中，$\sum$表示对一组商品项目中的所有项目求和。

2011年（$t=2011$）的普通汽车运行费用的未加权综合物价指数为

$$I_{2011} = \frac{3.52 + 6.25 + 145.00 + 1\,040.00}{1.30 + 2.10 + 130.00 + 820.00} \times 100 = \frac{1\,194.7}{953.40} \times 100 = 125.3$$

根据未加权综合物价指数，我们看到普通汽车运行费用的价格1990~2011年上涨了25.3%。但值得注意的是，汽车运行费用的未加权综合指数近似于一个复合物价指数，它受单价高的商品项目的影响严重。因此，像汽油、润滑油这些单价较低的商品项目，就受到像轮胎、保险费用这些单价较高的商品项目的制约。汽车运行费用的未加权综合物价指数就受轮胎和保险费用价格变动的影响严重。

由于未加权综合指数受一个或多个高价商品项目的影响，因此，这种形式的综合指数应用并不广泛。当商品项目的使用量不同时，加权综合物价指数给出了一个更好的对比方法。[⊖]

加权综合物价指数（weighted aggregate price index）的原理是对一组商品项目中的每一个商品项目，依据其重要性进行加权。在大多数情况下，商品的使用量是测量其重要性的最好尺度。因此，对一组商品项目组合中的每一个项目，需要有其使用量的资料。表20-4以一辆每年大约行驶15 000英里的具有代表性的中型汽车为依据，给出了各项汽车运行费用的年使用量信息。表中列出的数量权数表明该车行驶情况的年期望使用量。

表20-4 汽车运行费用指数的年使用量信息

商品项目	数量权数[①]
汽油（加仑）	1 000
润滑油（夸脱）	15
轮胎	2
保险费用	1

①汽车以每年行驶15 000英里为依据，轮胎的使用量以轮胎的寿命为30 000英里依据。

令Q_i表示第i个商品项目的使用量，则t期的加权综合物价指数公式为

$$I_t = \frac{\sum P_{it} Q_i}{\sum P_{i0} Q_i} \times 100 \qquad (20\text{-}3)$$

式中，$\sum$表示对一组中所有的商品项目求和。

应用到汽车运行费用问题，加权综合物价指数是2011年总运行费用与1990年的总运行费用的对比。

令$t=2011$，根据表20-4中的数量权数，我们可以得出2011年汽车运行费用的综合物价指数为

$$I_{2011} = \frac{3.52 \times 1\,000 + 6.25 \times 15 + 145.00 \times 2 + 1\,040.00 \times 1}{1.30 \times 1\,000 + 2.10 \times 15 + 130.00 \times 2 + 820.00 \times 1} \times 100 = \frac{4\,943.75}{2\,411.5} \times 100 = 205.0$$

从这个加权综合物价指数中，我们可以得到汽车运行费用的价格在1990～2011年上涨了105%。

显然，与未加权综合物价指数相比，加权综合物价指数更准确地反映了1990～2011年汽车运行费用的价格变动。考虑到汽油的使用量能抵消保险费用较小的增长速度，因此，在反映汽车运行费用增速方面，加权综合指数比未加权综合指数有较大增长。一般地，以使用量为权数的加权综合指数是比较适合于编制一组商品项目的价格指数。

在加权综合物价指数公式（20-3）中，我们注意到，数量Q_i没有显示时间的第2个下标，其原因是数量Q_i被看成是固定的，而且不像价格那样随时间的变动而变动。固定权数或数量是被指数设计者指定的，其水平被认为是具有代表性的使用量。一旦Q_i得到确定，它们在编制指数的各个时期内保持不变。计算2011年以外其他年份的指数，还需要搜集新的价格资料P_{it}，但权数Q_i保持不变。

以基期的使用量为权数是固定权数的加权综合指数的一个特殊情形。在这种情形下，$Q_i = Q_{i0}$，其中下标0表示基期。因此式（20-3）变为

$$I_t = \frac{\sum P_{it} Q_{i0}}{\sum P_{i0} Q_{i0}} \times 100 \qquad (20\text{-}4)$$

以基期数量为权数的加权综合指数被称为**拉斯贝尔指数**（Laspeyres Index）。

另一个确定权数的方法是每个时期修正数量权数。每年计算指数时所用的数量为Q_{it}，使用这些数量为权数的t期的加权综合物价指数为

$$I_t = \frac{\sum P_{it} Q_{it}}{\sum P_{i0} Q_{it}} \times 100 \qquad (20\text{-}5)$$

注意，对基期（0期）和t期用相同的数量加权。然而，若权数是以t期而不是以基期的使用量为依据，则加权综合指数被称为**派许指数**（Paasche Index）。以报告期的使用量为权数是其优点，但这种计算综合指数的方法也有两个缺点：一是通常使用量Q_{it}必须每年重新确定，因此，增加了搜集资料的时间和成本；二是每年必须重新计算前一年的指数，以反映新的数量权数的效果。正是由于这些缺点，拉斯贝尔指数应用更为广泛。汽车运行费用指数是以基期数量为权数计算的，因此它是拉斯贝尔指数。若使用2011年的使用量资料，我们可以得到派许指数。的确，由于汽车的节能性提高，汽油的使用量将会降低，因此派许指数就会不同于拉斯贝尔指数。

⊖ 如果每一商品项目的使用量相同，未加权综合指数相当于权数相等的加权综合指数。但在实践中，各个商品项目的使用量相同这种情况很少出现。

练 习

方法

2. 某个商品项目2011年的成本为10.75美元；以1994年为基期，价比为132。

   a. 17年来该商品项目成本上升或下降的百分比是多少？

   b. 1994年该商品项目的成本是多少？

应用

4. R&B饮料公司通过其设在艾奥瓦中部的零售网点，销售其生产的啤酒、葡萄酒及软饮料等全部产品。2008年和2011年的单价及2008年的销售数量（单位：箱）资料如下：

商品	2008年数量	单价（美元）	
	（箱）	2008年	2011年
啤酒	35 000	17.50	20.15
葡萄酒	5 000	100.00	118.00
软饮料	60 000	8.00	8.80

以2008年为基期，计算2011年R&B饮料销售的加权综合物价指数。

20.3 根据价比计算综合物价指数

在第20.1节我们给出了价比的定义，阐明如何根据报告期和基期的单价计算价比。现在，我们希望说明，如何像第20.2节编制综合物价指数那样，根据一组商品项目中每一个商品项目价比的信息，直接计算综合物价指数。由于未加权指数的局限，我们只讨论加权综合物价指数。我们仍以上一节的汽车运行费用指数为例，所需的四种费用的资料列在表20-5中。

表20-5 汽车运行费用指数的价比

商品项目	单价（美元）		价比	年用量
	1990（P_0）	2011（P_t）	（P_t/P_0）×100	
汽油（加仑）	1.30	3.52	270.8	1 000
润滑油（夸脱）	2.10	6.25	297.6	15
轮胎	130.00	145.00	111.5	2
保险费用	820.00	1 040.00	126.8	1

令 w_i 是适用于第 i 个商品项目价比的权数，则加权平均价格指数的一般形式为

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}} w_i}{\sum w_i} \times 100 \tag{20-6}$$

适当选择式（20-6）中的权数，将使我们能根据价比计算加权综合指数。适当选择使用量和基期价格的乘积为权数[⊖]，即

$$w_i = P_{i0} Q_i \tag{20-7}$$

将式（20-7）的权数 w_i 代入式（20-6）中，则加权平均价格指数可表示为

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}} P_{i0} Q_i}{\sum P_{i0} Q_i} \times 100 \tag{20-8}$$

对上式进行化简，可得加权平均价格指数的一个等价表达式为

$$I_t = \frac{\sum P_{it} Q_i}{\sum P_{i0} Q_i} \times 100$$

因此，我们看到以 $w_i = P_{i0} Q_i$ 为权数的加权平均价格指数，与第20.2节介绍的式（20-3）完全相同的加权综合物价指数。在式（20-7）中使用基期数量（即 $Q_i = Q_{i0}$），可以导出拉斯贝尔指数；在式（20-7）中使用报告期

⊖ 价格与数量的单位应该相一致，即如果给的是每箱的价格，则数量就应为箱数，而不应为其他的单位。

数量（即 $Q_i = Q_{it}$），可以导出派许指数。

让我们回到汽车运行费用的资料，利用表 20-5 中的价比和式（20-6），我们能计算出价比加权平均指数。利用式（20-7）的权数 $w_i = P_{i0}Q_i$，得到的计算结果列在表 20-6 中。指数 205 表明汽车运行费用增长了 105%。这与第 20.2 节计算的加权综合物价指数的增长幅度相同。

表 20-6 根据加权价比计算的汽车运行费用指数（1990 ~ 2011 年）

商品项目	价比 $(P_{it}/P_{i0}) \times 100$	基期价格（美元）P_{i0}	数量 Q_i	权数 $w_i = P_{i0}Q_i$	加权价比 $(P_{it}/P_{i0}) w_i \times 100$
汽油	270.8	1.30	1 000	1 300.00	352 040.00
润滑油	297.6	2.10	15	31.50	9 374.40
轮胎	107.7	130.00	2	260.00	28 990.00
保险费用	126.8	820.00	1	820.00	103 976.00
			合计	2 411.50	494 380.40

$$I_{2001} = \frac{494\,380.40}{2\,411.50} = 205$$

练习

方法

☆6. 下表是三个商品项目的价比、基期的价格和使用量，计算报告期三个商品项目的加权综合物价指数。

商品项目	价比	基期	
		价格	使用量
A	150	22.00	20
B	90	5.00	50
C	120	14.00	40

应用

8. 由四种股票构成的一个证券投资组合，它们的买入价、现价及股票数量资料如下表所示。

股票	买入价（美元/股）	现价（美元/股）	股票数量
Holiday Trans	15.50	17.00	500
NY Electric	18.50	20.25	200
KY Gas	26.75	26.00	500
PQ Soaps	42.25	45.50	300

根据该投资组合的资料编制加权平均价格指数，并解释这个指数。

20.4 一些重要的价格指数

我们已经介绍了用于计算单个商品项目或一组商品项目的价格指数。现在我们来考虑度量商业和经济状况的一些重要的价格指数。特别地，我们将考虑消费者价格指数、生产者价格指数和道琼斯平均指数。

20.4.1 消费者价格指数

消费者价格指数（consumer price index，CPI），由美国劳工统计局每个月定期公布，是美国生活费用水平的基本度量。[⊖]用于编制这个指数的商品项目组由 400 个商品项目构成一个购物篮，购物篮中包括食品、住房、服装、交通运输和医疗保健等商品项目。消费者价格指数[⊖]是固定权数的加权综合物价指数。通过对全美国城市家庭的定期调查，获得购物篮中每个商品项目的权数。

以 1982 ~ 1984 年的消费者价格指数为 100，2012 年 2 月的消费者价格指数为 227.7。这意味着以 1982 ~ 1984 年为基期，目前购物篮中的货物和服务的消费价格上升了 127.7%。图 20-1 是 1960 ~ 2010 年 50 年的消费者价格指数的时间数列，从该图形上可以看出，消费者价格指数反映出 20 世纪 70 年代后期到 80 年代初期是通货膨胀剧烈的时期。

⊖ 消费者价格指数中包括服务费（例如，医生和牙医的账单）以及与购买和使用每种商品直接相关的所有税费。

⊖ 劳工统计局实际上公布了两个消费者价格指数：一个是所有城市消费者的消费者价格指数（CPI-U），二是对城市工薪阶层和文职人员的修正的消费者价格指数（CPI-W）。其中城市消费者的消费者价格指数应用比较广泛，它由《华尔街日报》定期发布。

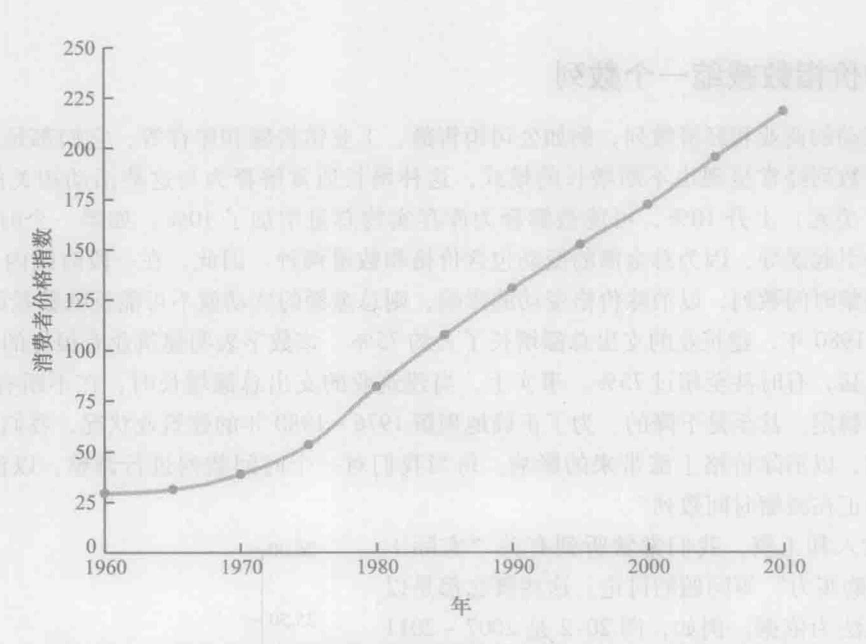


图 20-1 1960~2010年的消费者价格指数（基期1982~1984年=100）

20.4.2 生产者价格指数

生产者价格指数（producer price index，PPI），也是由美国劳工统计局每个月定期公布，它用于度量美国初级市场每个月价格的变动。[㊀]生产者价格指数依据每种产品在非零售市场上首次交易时的价格来计算，其中包括在市场上进行交易的所有日用品。这项统计调查涵盖了原材料、在每个加工阶段上制造和加工的产品，也包括制造业、农业、林业、渔业、采矿业和电力，以及公共事业等各行业的出口产品。这个指数一个常见的用途是作为消费者价格和生活费用未来趋势的一个领先指标。生产者价格指数的增长将导致生产价格上涨，而这种上涨通过较高的零售价格，最终传递给消费者。

生产者价格指数中各个商品项目权数的依据是货物的价值，是根据拉斯贝尔方法计算的加权平均价格指数。以1982年的生产者价格指数为100，2012年2月的生产者价格指数为201.6。

20.4.3 道琼斯平均指数

道琼斯平均指数（Dow Jones averages）[㊁]是用来反映普通股票价格趋势和波动变化的指数。最著名的道琼斯指数是道琼斯工业平均指数（DJIA），它是根据30家大公司的普通股票价格计算出来的。它是这些股票价格的总和与一个数相除得到的，这个数因为该指数公司拆股和股票转换而需要不断调整修订。它不像我们已学过的其他价格指数，道琼斯工业平均指数不能表示为基期价格的一个百分数。表20-7是2012年2月计算道琼斯工业平均指数所使用的30家具体公司。

表 20-7 道琼斯工业平均指数使用的30家公司（2012年2月）

3m	迪士尼	卡夫食品
美国铝业	杜邦	麦当劳
美国运通	埃克森美孚	默克制药
美国电话电报公司	通用电气	微软
美洲银行	惠普	辉瑞
波音	家得宝	宝洁
卡特彼勒	IBM	旅行者
雪佛龙	英特尔	联合科技
思科	强生	威瑞森
可口可乐	摩根大通	沃尔玛

道琼斯平均指数还有20种交通运输股票指数和15种公用事业股票指数。道琼斯平均指数每天都会被计算出来，并在《华尔街日报》和其他金融出版物上发布。

㊀ 生产者价格指数只是对本国产品价格变动的测度，进口产品不包括在内。

㊁ 查尔斯·亨利·道于1884年7月3日在"*Customer's Afternoon Letter*"上首次发表了他的股票平均价格指数。在最初的指数中包括11只股票，其中9只是铁路股票。1928年10月1日首次公布了可比的道琼斯工业平均指数。

20.5 根据物价指数减缩一个数列

许多随时间变动的商业和经济数列，例如公司销售额、工业销售额和库存等，它们都是以美元计量的。随着时间的推移，这些数列经常呈现出不断增长的模式，这种增长通常解释为与这些活动相关的实物量的增长。例如，库存总金额（美元）上升10%，可能被解释为库存实物总量增加了10%。如果一个时间数列以美元计量，这样的解释可能会引起误导，因为总金额的变动包含价格和数量两种。因此，在一段时间内，当价格的变动显著时，如果我们不调整时间数列，以消除价格变动的影响，则总金额的变动就不可能反映数量的变动。

例如，1976~1980年，建筑业的支出总额增长了大约75%。该数字表明建筑业有很好的增长态势。但是，建筑物的价格上升迅猛，有时甚至超过75%。事实上，当建筑业的支出总额增长时，在不断有新房开工的情况下，建筑业却保持相对稳定，甚至是下降的。为了正确地理解1976~1980年的建筑业状况，我们需要用一个价格指数来调整总支出数列，以消除价格上涨带来的影响。每当我们对一个时间数列进行调整，以消除价格上涨的影响时，我们说：我们正在减缩时间数列[㊀]。

对于个人的收入和工资，我们常常听到有关“实际工资”[㊁]或工资的“购买力”等问题的讨论。这些概念都是以减缩小时工资的想法为依据。例如，图20-2是2007~2011年电工的小时工资曲线，从图上看到工资有明显的上涨趋势，从每小时23.12美元增加到每小时25.44美元。对小时工资这样的上涨电工是否满意？应该根据他们工资的购买力是否增长来回答这个问题。如果我们可以对2007年每小时工资23.12美元的购买力与2011年每小时工资25.44美元的购买力进行比较，我们将能更好地判断工资的相对增加情况。

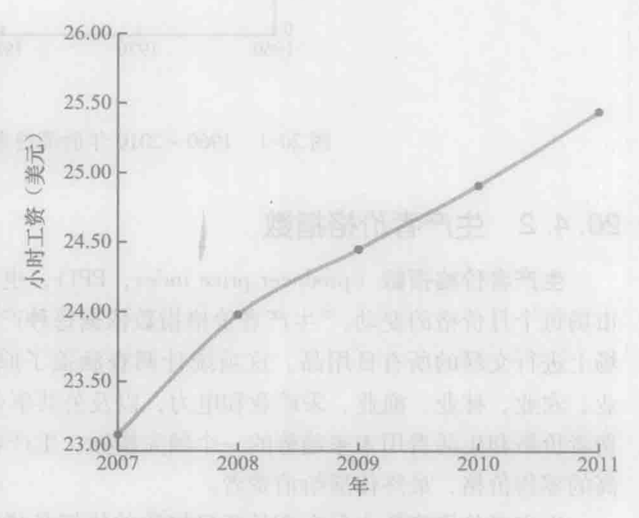


图20-2 电工现实的小时工资

表20-8列出了2007~2011年的小时工资率与消费者价格指数（以1982~1984年=100）。利用这些数据，我们将表明如何利用消费者价格指数来减缩小时工资指数。每年的小时工资率除以当年相应的消费者价格指数的数值，再乘以100，得到了减缩的数列。表20-9给出了电工减缩后的小时工资指数，图20-3显示了减缩的或实际的工资曲线。

表20-8 2007~2011年电工的小时工资和消费者价格指数

年份	小时工资（美元）	消费者价格指数（CPI）	年份	小时工资（美元）	消费者价格指数（CPI）
2007	23.12	207.3	2010	24.91	218.1
2008	23.98	215.3	2011	25.44	224.9
2009	24.45	214.5			

资料来源：美国劳工统计局。消费者价格指数以1982~1984年=100计算。

表20-9 2007~2011年电工小时工资的减缩数列

年份	减缩后的小时工资（美元）	年份	减缩后的小时工资（美元）
2007	(23.12/207.3)×100=11.15	2010	(24.91/218.1)×100=11.42
2008	(23.98/215.3)×100=11.14	2011	(25.44/224.9)×100=11.31
2009	(24.45/214.5)×100=11.40		

㊀ 对时间数列进行减缩可以消除通货膨胀的影响。

㊁ 实际工资比现实工资更能测量购买力。的确，许多劳工合同都要求工资依据生活费用的变动加以调整。

对于2007~2011年电工的实际工资或购买力，工资的减缩数列能说明什么问题呢？根据基期的价格（以1982~1984年=100），小时工资率在这段时间内几乎是平坦的。在消除了通货膨胀的影响之后，我们发现工人的购买力四年间仅增加了0.16美元，图20-3很清晰地反映出这一点。因此，用物价指数来减缩一个时间数列的优点是：我们对实际货币正在发生的变动有一个更清楚的描述。

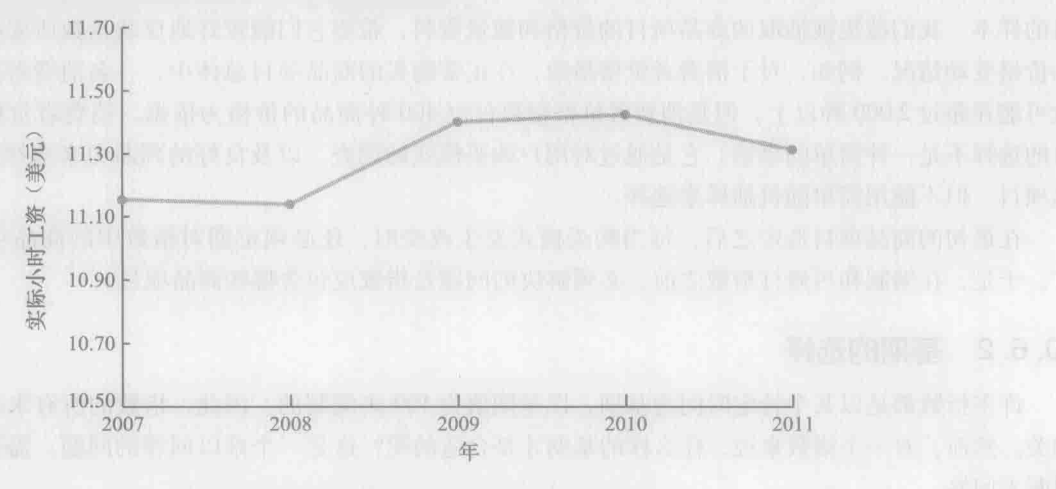


图20-3 2007~2011年电工实际的小时工资

一个随着时间推移的数列进行减缩的方法，在计算国内生产总值上有重要应用。国内生产总值是一个国家生产的全部产品和服务的总价值。显然，如果不用价格指数减缩国内生产总值，随着时间的推移，国内生产总值将会随着价格的上涨而增长。因此，要调整产品和服务的总价值，以反映生产和销售的产品、服务的实际数量的变动，国内生产总值就必须用一个价格指数来减缩计算。这一方法同实际工资的计算方法类似。

练习

应用

☆10. 注册护士每小时平均工资2007年为30.04美元，2011年为33.23美元。消费者价格指数2007年为207.3，2011年为224.9，回答下列问题。

a. 减缩注册护士的2007年与2011年的小时工资率，并求出实际工资率。

b. 对注册护士，2007~2011年现实的小时工资变动的百分比是多少？

c. 对注册护士，2007~2011年实际小时工资变动的百分比是多少？

12. 美国人口普查局追踪从制造业到零售业的产品总值。2009~2011年9月的计算机和电子产品的总产值以及这些月份的消费者价格指数和生产者价格指数如下表所示。

年	制造业的总产值（10亿美元）	消费者价格指数	生产者价格指数
2009	29.1	216.0	173.4
2010	33.3	218.4	180.2
2011	32.9	226.9	192.5

a. 用消费者价格指数减缩计算机和电子产品的总产值。

b. 用生产者价格指数减缩计算机和电子产品的总产值。

c. 你认为消费者价格指数和生产者价格指数哪一个更合适于减缩这些总产值？为什么？

20.6 物价指数：其他注意事项

在前面几节中，我们描述了几种计算物价指数的方法，讨论了一些重要的物价指数的应用，并且介绍了用物价指数来减缩时间数列的步骤。但为了加深我们对编制物价指数及其用途的了解，还必须考虑一些其他问题。这些问题将在本节进行讨论。

20.6.1 商品项目的选择

物价指数的首要目的在于测量指定的一类商品项目、产品等的价格随时间推移而发生的变动。由于同类商品的项目很多，所以不能以这一类中的所有商品项目为依据来编制指数。因此，必须选用一个具有代表性的商品项目的样本。我们搜集被抽取的商品项目的价格和数量资料，希望它们能较好地反映指数所要描述的所有商品项目的价格变动情况。例如，对于消费者价格指数，在正常购买的商品项目总体中，一名消费者要考虑的商品项目总数可能在超过2 000种以上，但是消费者价格指数仅以400种商品的价格为依据。消费者价格指数中具体商品项目的选择不是一件简单的事情。它是通过对用户购买模式的调查，以及良好的判断力来选择进入样本的400种商品项目，但不能用简单随机抽样来选择。

在最初的商品项目选定之后，每当购买模式发生改变时，还必须定期对指数中的商品项目进行再审查与修订。于是，在编制和再修订指数之前，必须解决的问题是指数应包含哪些商品项目。

20.6.2 基期的选择

许多指数都是以某个特定时间为基期，以基期值为100来编制的。因此，指数的所有未来值都与基期值息息相关。然而，对一个指数来说，什么样的基期才是合适的呢？这是一个难以回答的问题，需要根据指数编制者的判断来回答。

2011年美国政府编制的许多指数都以1982年为基期。作为一个普遍的准则，基期不应与报告期离得太远。例如，对大多数人来说，以1945年为基期计算的消费者价格指数是难以理解的，因为人们不熟悉1945年的生活情况。因此，许多指数的基期需要定期调整为较近的时期。在1988年，消费者价格指数的基期从1967年调整为1982～1984的平均数，生产者价格指数目前以1982年为基期（即以1982年=100）。

20.6.3 品质的改变

物价指数的目的是测量价格随时间的推移而发生的变动。理想的做法是搜集相同的商品项目在几个不同时期上的价格资料，然后计算指数。一个基本假设是在每一个时期，相同商品项目的价格是确定的。当一种产品因时间推移而发生品质改变时，就可能出现问题。例如，一个制造商年复一年地发生使用廉价的原料、减少产品的功能等问题，则可能会改变产品的品质。在接下来的年份里，该产品的价格可能上升，但这个价格所对应的产品品质却在下降。因此，某种商品项目的实际价格可能上涨得比定价显示得要高。对于一种品质下降的商品项目，调整指数即使是可能的，也是困难的。

品质的重大改变也能引起产品价格的上涨，因此在计算指数时，应该删除因品质改变而导致价格上涨的部分。但是，当一种商品项目因品质提高而导致价格上涨时，调整指数即使是可能的，也是困难的。

虽然，在编制物价指数时，通常的做法是忽视微小的品质改变，但是重大的品质改变必须考虑，因为随着时间的积累它们可以改变产品的性能。如果产品的性能发生了改变，指数就必须为此加以修订。在某些情况下，该产品可能会从指数的计算中删除。

但是，在某些情况下，伴随着品质的重大改变是价格的下降，这个极为典型的情形体现在20世纪90年代以及21世纪早期的个人电脑的案例上。

20.7 物量指数

除了前面几节所描述的物价指数以外，其他类型的指数也是有用的。特别地，指数的另一个应用是测定物量随时间的推移而发生的变动。这种类型的指数被称为**物量指数**（quantity index）。

回顾第20.2节中所讨论的加权综合物价指数，为了计算t期的指数，我们需要基期单价（P_0）和t期的单价（P_t）。式（20-3）给出的加权综合物价指数的公式为

$$I_t = \frac{\sum P_{it}Q_i}{\sum P_{i0}Q_i} \times 100$$

分子$\sum P_{it}Q_i$表示t期各个指数项目固定数量的总价值，分母$\sum P_{i0}Q_i$表示基期各个指数项目相同数量的总价值。

加权综合物量指数的计算与加权综合物价指数的计算类似。对每一种商品项目在基期和t期的数量进行测量，分别用Q_{i0}和Q_{it}表示第i种商品项目在基期和t期的数量。然后，对数量用一个固定的价格、增加值或其他因素进行加权。产品的“增加值”是销售额减去投入的购买成本，则t期的加权综合物量指数的计算公式为

$$I_t = \frac{\sum Q_{it}w_i}{\sum Q_{i0}w_i} \times 100 \tag{20-9}$$

在有些物量指数中，使用第i项商品项目基期的价格（P_{i0}）为权数，这种情形下的加权综合物量指数为

$$I_t = \frac{\sum Q_{it}P_{i0}}{\sum Q_{i0}P_{i0}} \times 100 \tag{20-10}$$

物量指数也可以用物量比的加权形式来计算，这种方法的物量指数公式的公式为

$$I_t = \frac{\sum \frac{Q_{it}}{Q_{i0}}Q_{i0}P_i}{\sum Q_{i0}P_i} \times 100 \tag{20-11}$$

这个公式是第20.3节中的加权价比指数公式（20-8）的数量表现形式。

由美国联邦储备局编制的**工业生产指数**（index of industrial production）可能是最著名的物量指数，它以2002年为基期，且每个月发布一次。编制该指数的目的是为了度量各类制造业生产水平的数量变动，但不包括采矿业和公用事业。2012年2月该指数为96.2。

练习

方法

☆14. 下面是三个商品项目1997年和2011年的销售量及1997年的销售价格数据，计算2011年的销售量的加权综合指数。

商品项目	销售量		单价（美元）
	1997年	2011年	1997
A	350	300	18.00
B	220	400	4.90
C	730	850	15.00

应用

16. 下表是一个汽车销售商1994年和2011年销售三种型号汽车的销售量数据，计算物量比，并利用物量比编制2011年的加权综合物量指数。

型号	销售量		1994年平均售价（美元）
	1994年	2011年	
Sedan	200	170	15 200
Sport	100	80	17 000
Wagon	75	60	16 800

小 结

在商业和经济环境中，价格指数和物量指数是价格和数量变动的重要测度。价比是某一商品项目报告期与基期单价之比再乘以100。如果价比为100，则说明报告期的价格与基期的价格没有差别。编制综合物价指数是对一组给定的商品项目或产品的全部价格变动进行综合度量。通常，一个加权综合物价指数中的商品项目被它们的使用量加权。加权综合物价指数也可以利用指数中商品项目的使用量，用对价比进行加权的方法来计算。

消费者价格指数和生产者价格指数是两个应用广泛的指数，它们分别以1982～1984年和1982年为基期。道琼斯工业平均价格指数是另一个应用广泛的物价指数，它是纽约股票交易所上市的30家大公司的普通股票价格的加权和。和许多其他指数不同，它不能表示为基期数值的百分数。

物价指数经常用于减缩一些随时间推移而变动的

经济数列。我们介绍了如何用消费者价格指数减缩每小时的工资，从而得到实际工资指数。选择指数中所包含的商品项目、选择指数的基期，以及因品质改变而进行的调整，是编制指数时重要的其他注意事项。物量指数的讨论较为简单扼要，工业生产指数是作为一个重要的物量指数而被提及。

关键术语

Price relative 价比 一个给定商品项目的一种物价指数，它用该商品项目报告期的单价与基期的单价对比再乘以100来计算。

Aggregate price index 综合物价指数 根据一组商品项目的价格而得出的一种复合物价指数。

Weighted aggregate price index 加权综合物价指数 对一组商品项目的价格，依据各自的重要性进行加权而得到的一种复合物价指数。

Laspeyres index 拉斯贝尔指数 以每一种商品项目的基期数量作为权数的加权综合物价指数。

Paasche index 派许指数 以每一种商品项目的报告期数量作为权数的加权综合物价指数。

Consumer price index (CPI) 消费者价格指数 每个月定期发布的一种物价指数，它利用购物篮中的商品和服务的价格变动来度量消费价格随时间的推移而发生的变动。

Producer price index 生产者价格指数 每月发布一次的一种物价指数，旨在度量初级市场上销售的商品（即在非零售市场上首次购买某种商品）的价格变动的情况。

Dow Jones averages 道琼斯平均指数 是用来显示在纽约证券交易所上市的普通股票价格趋势与波动变化的一种综合物价指数。

Quantity index 物量指数 旨在度量物量随时间的推移而变动的一种指数。

Index of industrial production 工业生产指数 旨在度量工业产品实物量或工业生产水平随着时间的推移而变动的一种物量指数。

重要公式

t 期的价比

$$t\text{ 期的价比} = \frac{t\text{ 期的价格}}{\text{基期的价格}} \times 100 \qquad (20\text{-}1)$$

t 期的未加权综合物价指数

$$I_t = \frac{\sum P_{it}}{\sum P_{i0}} \times 100 \qquad (20\text{-}2)$$

t 期的加权综合物价指数

$$I_t = \frac{\sum P_{it} Q_i}{\sum P_{i0} Q_i} \times 100 \qquad (20\text{-}3)$$

加权平均物价指数

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}} w_i}{\sum w_i} \times 100 \qquad (20\text{-}6)$$

式（20-6）的权数

$$w_i = P_{i0} Q_i \qquad (20\text{-}7)$$

加权综合物量指数

$$I_t = \frac{\sum Q_{it} w_i}{\sum Q_{i0} w_i} \times 100 \qquad (20\text{-}9)$$

补充练习

18. Nickerson 制造公司的四种产品的运输数量和费用资料如下：

产品	基期数量 2005 年	单位平均运费（美元） 2005 年	2011 年
A	2 000	10.50	15.90
B	5 000	16.25	32.00
C	6 500	12.20	17.40
D	2 500	20.00	35.50

a. 计算每种产品的价比。

b. 计算反映四年来运输费用变动的加权综合物价指数。

20. Boran 证券经纪人有限公司选择四只股票来编制反映股票市场行情的指数。以 2009 年为基期，四只股票 2011 年 1 月及 3 月的每股价格如下表所示。基期的数量以四只股票的历史数量为基础。

股票	行业	2009 年数量	每股价格（美元） 2009 基期	2011 年 1 月	2011 年 3 月
A	石油	100	31.50	22.75	22.50
B	计算机	150	65.00	49.00	47.50
C	钢铁	75	40.00	32.00	29.50
D	房地产	50	18.00	6.50	3.75

以2009年为基期，计算2011年1月和3月的Boran指数，该指数告诉你有关股票市场生产的情况，对此有何评论。

22. 假设在2001年平均一个男用剃须刀需购买一个剃刀手柄和使用17个刀片，2001~2011年的价比如下表所示。根据2011年加权价比编制男用剃须刀费用指数。

商品项目	2001平均使用量	基期价格（美元）	2001~2011年价比
剃须柄	1	7.46	126.9
刀片	17	1.90	153.7

24. 精算师是精通风险数学的分析师。精算师常常受雇于保险公司，并负责设置保险费用。下表是2008~2011年精算师的中位数年薪，同时给出了每一年的消费者价格指数（U. S. Census Bureau）。用消费者价格指数减缩年薪数据为不变价，以不变价评论年薪的变化。

年份	中位数年薪（1 000美元）	消费价格指数
2008	84 810	215.3
2009	87 210	214.5
2010	87 650	218.1
2011	91 060	224.9

26. 一个大型制造公司2007年和2011年的产量和产值资料如下表所示。根据这些资料，计算加权综合物量指数。评论该物量指数的含义。

产品	数量		产值（美元）
	2007年	2011年	
A	800	1 200	30.00
B	600	500	20.00
C	200	500	25.00

Statistics for Business and Economics

(12th Edition)

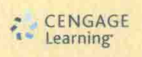
商务与经济统计（原书第12版）

这是一本广泛流行于美国高校并被誉为最经典的商务统计教材！该书在中国引进十多年间，一直被北京大学、清华大学、中国人民大学、复旦大学、上海交通大学、南开大学和中山大学等众多高校采用，深受师生推崇和喜爱。

应用性强是本书的最大特色。第12版在保留了以前版本的叙述风格和可读性的基础上，对内容进行了一定的修订，对个别章节做了更为合理的调整，并更新了一定数量的习题。作者精心设计了“方法”“应用”和“补充练习”三种题型，并设计了起提示、总结和建议作用的“注释和评论”，这些都体现出本书的实用特点。

本书适合本科生、MBA和研究生等多层次学者学习，也可作为从事工商行政管理和经济分析的各类人员手边的工具参考书。





www.cengageasia.com

投稿热线：(010) 88379007
客服热线：(010) 88379210 88361066
购书热线：(010) 68326294 88379649 68995259

华章网站：www.hzbook.com
网上购书：www.china-pub.com
数字阅读：www.hzmedia.com.cn







定价：109.00元（附光盘）